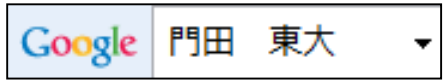
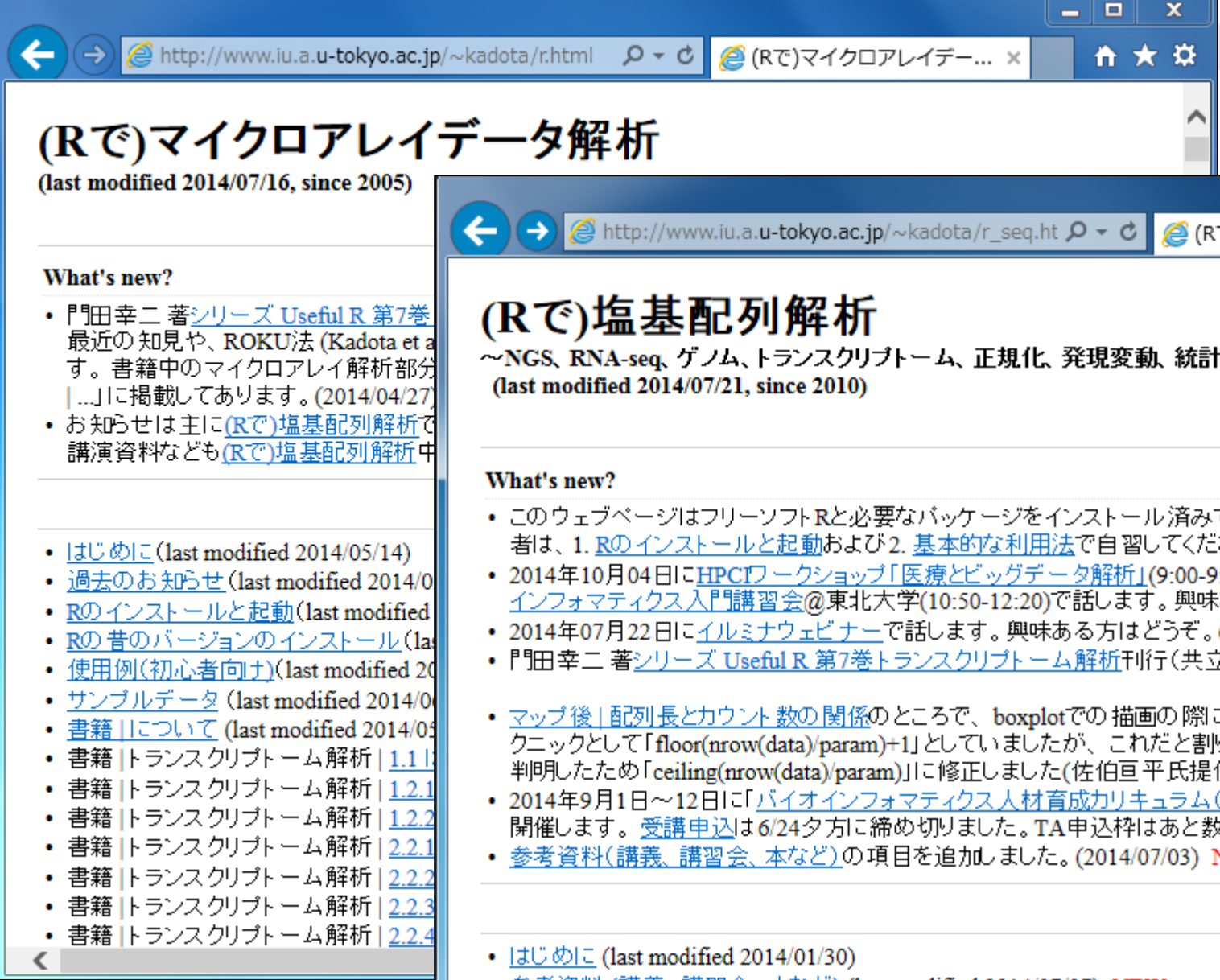
目的：2011年秋に2回行ったウェビナー以降のアップデート情報提供

トランスクリプトーム データ解析戦略2014

東京大学・大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
門田幸二（かどた　こうじ）
kadota@iu.a.u-tokyo.ac.jp
http://www.iu.a.u-tokyo.ac.jp/~kadota/





これやっているヒトです

Google ngs 解析

ウェブ 画像 ニュース 動画 ショッ

約 90,400 件 (0.12 秒)

NGS解析 | タカラバイオ株式会社

catalog.takara-bio.co.jp › TOP › ウェ

NGS解析 ... PCR(核酸増幅・ライブラ
メチル化**解析**. MethylEasy™ Xceed
EpiTaq™ HS (for bisulfite-treated D

(Rで)塩基配列解析

www.iu.a.u-tokyo.ac.jp/~kadota/r_se

このページは、次世代シーケンサー(N
短い塩基配列(short read)データ**解析**
あり、特にアグリバイオインフォマティ

次世代シークエンサ(NGS)解

d.hatena.ne.jp/sesejun/20100521/p1

2010/05/21 - 次世代シークエンサ (NG
2010/5/25. ... 既に**解析**をガシガシや
か, そんなん使わん, とか突っ込み歓

よく分かる次世代シークエン

chromatin.med.kyushu-u.ac.jp/arc

大学院講義でも利用しており、情報更新は頻繁に行っています。**少人数のスタッフで100人規模の実験系受講生を私が心穏やかに教えられる**ようにしています。内容はR中心ですが、NGS関連キーワードでもよく引っかかるようです。「はじめに」のところでも書いてありますが…

(Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル、バイオインフォマティクス～

(last modified 2014/07/21, since 2010)

What's new?

- このウェブページはフリーソフトRと必要なパッケージをインストール済みである前提で記述しています。初心者は、1. Rのインストールと起動および2. 基本的な利用法で自習してください。(2014/07/21) NEW
- 2014年10月04日にHPCIワークショップ「医療とビッグデータ解析」(9:00-9:20)に引き続いて 中級者向けバイオインフォマティクス入門講習会@東北大学(10:50-12:20)で話します。興味ある方はどうぞ。(2014/07/16) NEW
- 2014年07月22日にイルミナウェビナーで話します。興味ある方はどうぞ。(2014/06/30) NEW
- 門田幸二 著シリーズ Useful R 第7巻トランスクリプトーム解析刊行(共立出版)

- マップ後 | 配列長とカウント数の関係のところで、boxplotでの描画の際にparam個で分割(20分割など)するテクニックとして「floor(nrow(data)/param)+1」としていましたが、これだと割り切れる場合でも+1してしまうことが判明したため「ceiling(nrow(data)/param)」に修正しました(佐伯亘平氏提供情報)。(2014/07/03) NEW
- 2014年9月1日～12日に「バイオインフォマティクス人材育成カリキュラム(次世代シークエンサ)速習コース」を開催します。受講申込は6/24夕方に締め切りました。TA申込枠はあと数名です。(2014/07/21) NEW
- 参考資料(講義、講習会、本など)の項目を追加しました。(2014/07/03) NEW

- はじめに (last modified 2014/01/30)
- 参考資料(講義、講習会、本など) (last modified 2014/07/07) NEW
- 過去のお知らせ (last modified 2014/06/30) NEW
- Rのインストールと起動 (last modified 2014/07/07) NEW
- 基本的な利用法 (last modified 2014/07/20) NEW
- サンプルデータ (last modified 2014/07/17) NEW

トップページへ

- バイオインフォマティクス人材育成カリキュラム(次世代シークエンサ) | 速習コース (last modified 2014/07/17)

+ サイトマップ + English

東京大学大学院農学生命科学研究科

アグリバイオインフォマティクス教育研究ユニット

Agricultural Bioinformatics Research Unit

受講生の方へ　研究者の方へ

+ ホーム
+ 本ユニットについて
+ メンバー
+ 教育プログラム
+ 研究フォーラム
+ イベント
+ お問い合わせ
+ リンク

+ モバイルサイト

東京大学 THE UNIVERSITY OF TOKYO

ホーム > 教育プログラム > 各講義のページ

各講義のページ

（科目名をクリックすると各講義のページに移動します）

先端トピックス（セミナー・討論形式研究指導）
- 農学生命情報科学特別演習
- 農学生命情報科学特論 I
- 農学生命情報科学特論 II
- 農学生命情報科学特論 III
- 農学生命情報科学特論 IV

方法論（講義・実習を一体化）
- 生物配列統計学
- システム生物学概論
- 知識情報処理論
- オーム情報解析
- 機能ゲノム学
- 分子モデリングと分子シミュレーション

基礎（講義・実習を一体化）
- ゲノム情報解析基礎
- 構造バイオインフォマティクス基礎
- 生物配列解析基礎
- バイオスタティスティクス基礎論

カテゴリー	科目名	学期・単位	実施曜日
基礎	1. 生物配列解析基礎		
	生命科学のためのデータベースの利用と基本的な解析手法について講義します。データベースの基礎、配列データベース、機能データベース、ホモロジー検索、モチーフ解析などの基本的な手法について解説します。	夏・1	火曜
	2. ゲノム情報解析基礎		

講義風景(平成26年度)

この規模でLinuxを最初から教えるのはアリエマセン。ウェブツール系もかなり厳しいです。

Google トランスクリプトーム 解析

ウェブ 画像 ニュース 動画 ショッピング もっと見る 検索ツール

約 42,000 件 (0.22 秒)

他のキーワード: トランスクリプトーム解析方法 トランスクリプトーム解析 受託
トランスクリプトーム解析 次世代シーケンサー トランスクリプトーム解析 価格
トランスクリプトームとは

トランスクリプトーム - Wikipedia
ja.wikipedia.org/wiki/トランスクリプトーム
トランスクリプトミクス(transcriptomics)とは**トランスクリプトーム**を扱う学問である。**トランスクリプトーム**の様態は、例えばDNAマイクロアレイの様に一度に幾万ものmRNAを識別する能力を持つ技術をもって**解析**される。遺伝子産物であるmRNAの階層の要素 ...

[PDF]トランスクリプトーム解析・プロテオーム解析入門
www.jst.go.jp/nbdc/bird/jinzai/literacy/streaming/h22_2_1.pdf
目的 シークエンス**解析**によってDNA配列上で遺伝子と推. 定された部分について細胞レベルでmRNA量を測. 定・**解析**し・・・. 生体細胞内における遺伝子の発現状況を網羅的に把. 握することを目的としている。**トランスクリプトーム解析**とは? (1/3). トランスクリプ ...

[PDF]トランスクリプトーム解析の今昔 - アグリバイオインフォマティ...
www.iu.a.u-tokyo.ac.jp/~kadota/20110908_kadota.pdf
Sep 8 2011. 1.**トランスクリプトーム解析**の今昔. なぜマイクロアレイ? なぜRNA-Seq? 東京大学大学院農学生命科学研究科. アグリバイオインフォマティクス教育研究ユニット. 門田幸二(かどたこうじ) http://www.iu.a.u-tokyo.ac.jp/~kadota/ kadota@iu.

Amazon.co.jp: トランスクリプトーム解析 (シリーズ Useful R 7 ...
www.amazon.co.jp › 本 › 科学・テクノロジー › 数学
Amazon.co.jp: **トランスクリプトーム解析** (シリーズ Useful R 7): 門田 幸二, 金 明哲: 本.

Googleのリストアップアルゴリズムはよく分かりませんが、2011年秋のイルミナウェビナーシリーズ第1回で話した古い内容が上位にランクインしている状況はいかがなものかと…

Contents

Googleのリストアップアルゴリズムはよく分かりませんが、2011年秋のイルミナウェビナーシリーズ第1回で話した古い内容が上位にランクインしている状況はいかがなものかと…

- **情報収集先**
  - **イルミナのウェビナー**
  - (Rで)塩基配列解析
- トランスクリプトーム解析技術の進展と展望
  - Wet側
    - マイクロアレイ:3’発現アレイ → エクソンアレイ → トランスクリプトームアレイ
    - RNA-seq:Illumina short-readとPacBio long-read
  - Dry側
    - 遺伝子構造推定(ゲノム配列を利用)
    - データ解析手段(ウェブツール、Linux、R)
    - 転写物の発現量推定(トランスクリプトーム配列を利用)
    - 発現変動解析(Rパッケージを利用)

目的:2011年秋に2回行ったウェビナー以降のアップデート情報提供

illumina

アプリケーション　システム　インフォマティクス　臨床研究　受託サービス　サイエンス　サポート　カンパ

学会 & イベント / ウェビナー

イルミナでは、製品をご活用いただいている研究者による解析手法や研究事例、イルミナスタッフによる最新の製品や技術概
にご活用いただいているユーザーの方やこれからご利用をお考えの方におすすめな、サービス・サポート部による技術、機能、
ナーを開催しています。

どちらのウェビナーも開催日にご参加いただくことで、講師の先生やス
ご覧いただくことができます。ぜひご活用ください。

注)誠に恐れ入りますが、ウェビナーへのご参加およびサポートウェビナ
（イルミナウェビナーの録画は閲覧可能です）推奨のOSおよびブラウ

イルミナウェビナー

製品をご活用いただいている研究者による解析
事例、およびイルミナスタッフによる最新の製品
ご紹介します。

イルミナウェビナーの一覧を表示

Googleのリストアップアルゴリズムはよく分かりませんが、2011年秋のイルミナウェビナーシリーズ第1回で話した古い内容が上位にランクインしている状況はいかがなものかと…

2011/11/17　RNA-Seqをはじめよう！シリーズ
「Session 3 - データ解析のリテラシー」
東京大学大学院 農学生命科学研究科
門田 幸二 先生

第3回のセッションは、東京大学の門田先生によるデータ解析です。

(1) R...

続きを開く

2011/10/13　RNA-Seqをはじめよう！シリーズ
「Session 2 - RNA-Seq実験ノート：リード長とリード数のデザインとウェット実験の注意点」
東京大学大学院 新領域創成科学研究科
鈴木 穣 先生

第2回のセッションは、東京大学の鈴木先生をお迎えし

続きを開く

本ウェビナーのタイトル通り、2011年秋以降の更新情報提供が目的。

2011/09/08　RNA-Seq をはじめよう！シリーズ
「Session 1 -トランスクリプトーム解析の今昔：なぜマイクロアレイ？なぜRNAシーケンス？」
東京大学大学院 農学生命科学研究科
門田 幸二 先生

遺伝子発現解析はこれまでマイクロアレイが使われてきました。しかし、次世代シーケンサー技術が急速に広がり、次第にシーケンサーを使った解析に手法が移行しています。第1回のセッションでは、これまでの手法と比べてRNAシーケンスはどう違うのかについて、東京大学の門田先生にお話いただきます。

閉じる

http://www.illuminakk.co.jp/events/webinar_japan.ilmn

イルミナのウェビナー情報をフル活用

2011/11/17 RNA-Seqをはじめよう！シリーズ
「Session 3 - データ解析のリテラシー」
東京大学大学院 農学生命科学研究科
門田 幸二 先生

第3回のセッションは、東京大学の門田先生によるデータ解析です。

(1) R...

続きを開く

2011/10/13 RNA-Seqをはじめよう！シリーズ
「Session 2 - RNA-Seq実験ノート：リード長とリード数のデザインとウェット
東京大学大学院 新領域創成科学研究科
鈴木 穣 先生

第2回のセッションは、東京大学の鈴木先生をお迎えし、実際に実験を開

続きを開く

2011/09/08 RNA-Seq をはじめよう！シリーズ
「Session 1 -トランスクリプトーム解析の今昔：なぜマイクロアレイ？なぜRNAシーケンス？」
東京大学大学院 農学生命科学研究科
門田 幸二 先生

遺伝子発現解析はこれまでマイクロアレイが使われてきました。しかし、次世代シーケンサー技術が急速に広がり、次第にシーケンサーを使った解析に手法が移行しています。第1回のセッションでは、これまでの手法と比べてRNAシーケンスはどう違うのかについて、東京大学の門田先生にお話いただきます。

閉じる

2014/07/22 16:00-17:00 RNA-Seqシリーズ
「トランスクリプトームデータ解析戦略2014」
東京大学・大学院農学生命科学研究科・アグリバイオインフォマティクス教育研究ユニット
門田 幸二 先生

Illumina HiSeqシリーズをはじめとしたトランスクリプトーム解析用機器の技術革新は...

続きを開く

2014/06/24 RNA-Seqシリーズ
「進化するRNA-Seq：臨床検体からシングルセル解析まで -- ウェット・ドライ解析の実験ノート」
東京大学大学院新領域創成科学研究科　情報生命科学専攻
鈴木 穣　教授

2011年の秋に弊社ウェビナーでご講演いただきました東京大学大学院 鈴木 穣先生を再...

続きを開く

本ウェビナーのタイトル通り、2011年秋以降の更新情報提供が目的。

ウェビナーシリーズ第2回の鈴木穣先生の分は2014年6月にアップデートされているのでそれ以外の情報を紹介。

http://www.illuminakk.co.jp/events/webinar_japan.ilmn

イルミナのウェビナー情報をフル活用

2011/11/17 RNA-Seqをはじめよう！シリーズ
「Session 3 - データ解析のリテラシー」
東京大学大学院
門田 幸二 先生
第3回のセッション
(1) R...
続きを開く

2011/10/13 RNA-Seqをはじ
「Session 2 -
東京大学大学院
鈴木 穣 先生
第2回のセッション
続きを開く

2011/09/08 RNA-Seq をは
「Session 1 -
東京大学大学院
門田 幸二 先生
遺伝子発現解析
広がり、次第に
手法と比べてR
閉じる

2013/06/14 RNA-Seq をはじめよう！シリーズ
「0.1 pg の mRNA をシーケンスする
独立行政法人 理化学研究所 情報基盤セ
二階堂 愛 先生、笹川 洋平 先生
近年、同じ組織や培養環境にある細胞
続きを開く

2013/05/08 de novoシリーズ
「DDBJパイプラインによるRNA-seq配列のde novoアセンブル」
国立遺伝学研究所 大量遺伝情報研究室
中村 保一 先生、長崎 英樹 先生、谷沢 靖洋 先生
遺伝研で解析パイプラインについて、またパイプライン内部のス
続きを開く

2013/04/12 de novoシリーズ
「Miseqでのde novo実験」
沖縄科学技術大学院大学 DNAシーケンシング セクション
藤江 学 氏、小柳 亮 氏
De novo 解析にデスクトップ型次世代シーケンサー MiSeq を既に使っている方、これか...
続きを開く

2013/04/03 de novoシリーズ
「非モデル生物のRNA-seq解析 -- 実
基礎生物学研究所 生物機能解析センター
重信 秀治 先生
次世代DNAシークエンサーによって、
続きを開く

2014/07/22 16:00-17:00 RNA-Seqシリーズ
「トランスクリプトームデータ解析戦略2014」
東京大学・大学院農学生命科学研究科・アグリバイオインフォマティクス教育研究ユニット
門田 幸二 先生
Illumina HiSeqシリーズをはじめとしたトランスクリプトーム解析用機器の技術革新は...
続きを

2013/10/17 イルミナウェビナー
「NGSデータ解析プラットフォームMaser」
国立遺伝学研究所 生命情報研究センター 遺伝情報分析研究室(セルイノベーション)
藤井 信之 先生、吉武 和敏 先生
生命科学におけるNGSデータの利用は年々その利用頻度が増えています。
今回紹...
続きを開く

2013年のRNA-seq関連ウェビナーもアップデート情報といえる。例えば2011/11/17に紹介したDDBJ Read Annotation Pipelineは、中村保一先生らによる2013/05/08のウェビナーがあり。

本ウェビナーのタイトル通り、2011年秋以降の更新情報提供が目的。

http://www.illuminakk.co.jp/events/webinar_japan.ilmn

イルミナのウェビナー情報をフル活用

2011/11/17 RNA-Seqをはじめよう！シリーズ
「Session 3 - データ解析のリテラシー」
東京大学大学院
門田 幸二 先生

第3回のセッション

(1) R...

続きを開く

2011/10/13 RNA-Seqをはじ
「Session 2 -
東京大学大学院
鈴木 穣 先生

第2回のセッション

続きを開く

2011/09/08 RNA-Seq をは
「Session 1 -
東京大学大学院
門田 幸二 先生

遺伝子発現解析
広がり、次第に
手法と比べてR

閉じる

2013/06/14 RNA-Seq をはじめよう！シリーズ
「0.1 pg の mRNA をシーケンスする
独立行政法人 理化学研究所 情報基盤セ
二階堂 愛 先生、笹川 洋平 先生

近年、同じ組織や培養環境にある細胞

続きを開く

2013/05/08 de novoシリーズ
「DDBJパイプラインによるRNA-seq配列のde novoアセンブル」
国立遺伝学研究所 大量遺伝情報研究室
中村 保一 先生、長崎 英樹 先生、谷沢 靖洋 先生

遺伝研で解析パイプラインについて、またパイプライン内部のスパコン利用、DDBJパイ...

続きを開く

2013/04/12 de novoシリーズ
「Miseqでのde novo実験」
沖縄科学技術大学院大学 DNAシーケン
藤江 学 氏、小柳 亮 氏

De novo 解析にデスクトップ型次世代

続きを開く

2013/04/03 de novoシリーズ
「非モデル生物のRNA-seq解析 -- 実
基礎生物学研究所 生物機能解析センター
重信 秀治 先生

次世代DNAシークエンサーによって、

続きを開く

2014/07/22 16:00-17:00 RNA-Seqシリーズ
「トランスクリプトームデータ解析戦略2014」
東京大学・大学院農学生命科学研究科・アグリバイオインフォマティクス教育研究ユニット
門田 幸二 先生

Illumina HiSeqシリーズをはじめとしたトランスクリプトーム解析用機器の技術革新は...

続きを開く

本題と直接的な関係はありませんが、大変分かりやすい統計の話もあり。

2013/10/30 クリニカルシーケンスシリーズ
「臨床シーケンス解析をする上で理解しておきたい初歩からの遺伝統計学」
スタージェン株式会社
鎌谷 直之 先生

次世代シーケンサー（NGS）はゲノム研究に欠かせないツールとなりましたが、中でも診...

続きを開く

2013/10/17 イルミナウェビナー
「NGSデータ解析プラットフォームMaser」
国立遺伝学研究所 生命情報研究センター 遺伝情報分析研究室(セルイノベーション)
藤井 信之 先生、吉武 和敏 先生

生命科学におけるNGSデータの利用は年々その利用頻度が増えています。
今回紹...

続きを開く

お詫びと訂正

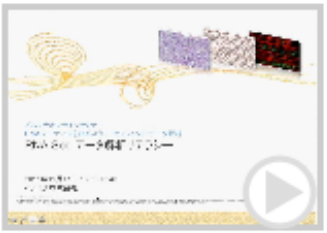
2011/11/17 RNA-Seqをはじめよう！シリーズ

「Session 3 - データ解析のリテラシー」
東京大学大学院 農学生命科学研究科
門田 幸二 先生

第3回のセッションは、東京大学の門田先生によるデータ解析です。

(1) R...

続きを開く



2011/10/13 RNA-Seqをはじめよう！シリーズ

「Session 2 - RNA-Seq実験ノート：リード長とリード数のデザイン
東京大学大学院 新領域創成科学研究科
鈴木 穣 先生

第2回のセッションは、東京大学の鈴木先生をお迎えし、実際に

続きを開く

2011/09/08 RNA-Seq をはじめよう！シリーズ

「Session 1 -トランスクリプトーム解析の今昔：なぜマイクロアレイ
東京大学大学院 農学生命科学研究科
門田 幸二 先生

遺伝子発現解析はこれまでマイクロアレイが使われてきました

Q & A

- Q：なぜsra.lite形式で配布するんですか？
  - A：ファイルサイズを大幅に圧縮できるからです
    - SRR002324.lite.sraファイル：約0.9GB
    - SRR002324.fastqファイル：約3.8GB
- Q：Linuxが使えないとだめ…ってことですよね?!
  - A：（今のところ）そう…ですね。…しかも…それ以外の様々な局面でLinux環境での作業が必要…

NGS解析はLinux上で行うのが基本

Nov 17 2011 11

イルミナウェビナーのウェブサイトから得られるPDFファイル中には間違いあり！スライド11の赤枠部分です。もちろん**悪意のないミス**です（爆）

お詫びと訂正

門田のウェブサイトで提供しているPDFファイルのほうは修正済みです。

Q & A

- Q:なぜsra.lite形式で配布するんですか？
  - A:ファイルサイズを大幅に圧縮できるからです
    - SRR002324.lite.sraファイル:約0.9GB
    - SRR002324.fastqファイル:約3.8GB
- Q:Linuxが使えないとだめ…ってことですよね?!
  - A:(今のところ)そう…ですね。…しかも…それ以外の様々な局面でLinux環境での作業が必要…

NGS解析はLinux上で行うのが基本

Nov 17 2011 11

Q & A

間違いが残ったままになっていたので変更しました(20130806)

- Q:なぜsra.lite形式で配布するんですか？
  - A:よくわかりません
- Q:Linuxが使えないとだめ…ってことですよね?!
  - A:(今のところ)そう…ですね。…しかも…それ以外の様々な局面でLinux環境での作業が必要…

NGS解析はLinux上で行うのが基本

Nov 17 2011 11

お詫びと訂正

Q & A

間違いが残ったままになっていたので変更しました(20130806)

- Q:なぜsra.lite形式で配布するんですか？
  - A:よくわかりません
- Q:Linuxが使えないと
  - A:(今のところ)そう..
    の様々な局面でLinux

NGS解析は

Nov 17 2011

http://rgm22.nig.ac.jp/mediawiki-ogareport/index.php/RAW_DATA_archiving/sharing_at_DDBJ

様々なファイル形式…

前頁のお詫びに講義資料で用いていた別のスライドを追加しました(20130806)

- 情報量:SRA-full > SRA-lite > FASTQ (> FASTA)
  - SRA-full:塩基配列、クオリティ情報、Intensity情報など画像以外の全て
  - SRA-lite:SRA-fullからIntensity情報を除いて軽量化したもの
  - FASTQ:塩基配列とクオリティ情報のみからなるもの
  - (FASTA:塩基配列のみからなるもの)
  - ファイルサイズ (SRA-full : SRA-lite : FASTQ : FASTA)
    - 6 : 3 : 2 : 1
    - 例:SRA-fullはFASTQの約3倍

FASTQ形式ファイルの利用が基本

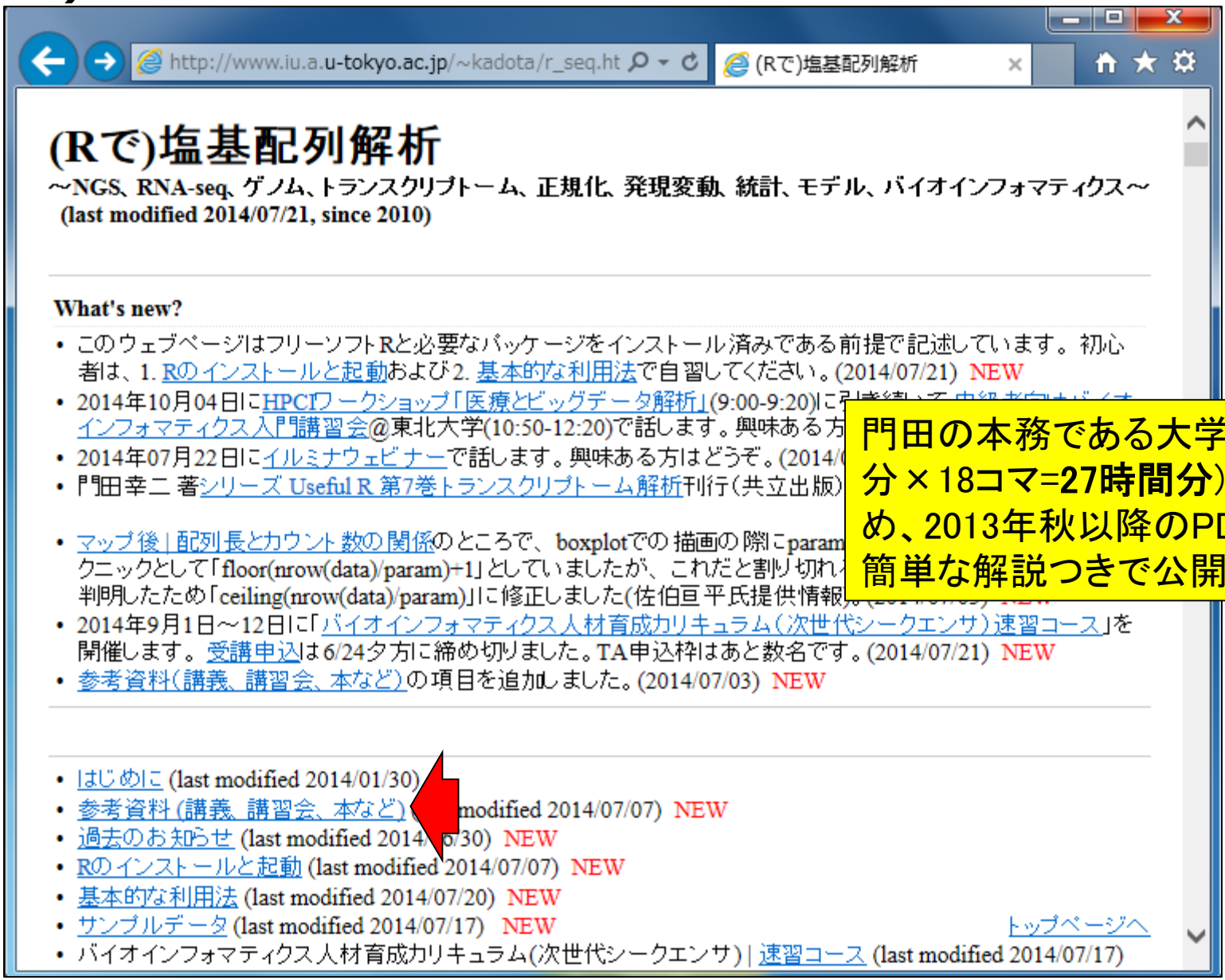
Jun 27 2013 12

門田提供のPDFファイルで1年以上前のものは参考程度にしてください。

Contents

- 情報収集先
  - イルミナのウェビナー
  - **(Rで)塩基配列解析**
- トランスクリプトーム解析技術の進展と展望
  - Wet側
    - マイクロアレイ：3’発現アレイ → エクソンアレイ → トランスクリプトームアレイ
    - RNA-seq：Illumina short-readとPacBio long-read
  - Dry側
    - 遺伝子構造推定（ゲノム配列を利用）
    - データ解析手段（ウェブツール、Linux、R）
    - 転写物の発現量推定（トランスクリプトーム配列を利用）
    - 発現変動解析（Rパッケージを利用）

(Rで)塩基配列解析の情報をフル活用



門田の本務である大学院講義(90分×18コマ=**27時間分**)スライドを含め、2013年秋以降のPDFファイルを簡単な解説つきで公開しています。

(Rで)塩基配列解析の情報をフル活用

- はじめに (last modified 2014/01/30)
- 参考資料(講義、講習会、本など) (last modified 2014/07/07) NEW
- 過去のお知らせ (last modified 2014/06/30) NEW
- Rのインストールと
- サンプルデータ (las
- バイオインフォマテ
- 書籍|トランスクリプ
- 書籍|トランスクリプ
- 書籍|トランスクリプ
- 書籍|トランスクリプ
- 書籍|トランスクリプ
- 書籍|トランスクリプ
- 書籍|トランスクリプ
- 書籍|トランスクリプ
- 書籍|トランスクリプ

参考資料(講義、講習会、本など) NEW

基本的に私門田の個人ページに記載してあるものです。かなり古い講演資料などの情報をもとに勉強されている方もいらっしゃるようですので、ここでは2013年秋以降の情報を載せておくとともに、大まかな内容についても述べておきます。講演予定のものについては、資料のアップは講演当日が基本です。50-100MB程度ありますがオリジナルのPowerPointファイルがほしい方はお気軽にリクエストしてください。講義資料としての利用などは事前連絡や謝辞も気にせずご自由にお使いください。

書籍

- **門田幸二**著(金明哲 編), シリーズ Useful R 第7巻トランスクリプトーム解析, 共立出版, 2014. ISBN: 978-4-320-12370-0
  内容:マイクロアレイとRNA-seq解析を例としてRを用いてトランスクリプトーム解析を行うための体系的な本としてまとめました。数式が苦手なヒト向けに、重みつき平均の具体的な計算例などを挙げてオプショ
  構成にしてあります。書籍中のRコードは「書籍|トランスクリプトーム解析|
- **門田幸二**,「トランスクリプトミクスの推奨データ解析ガイドライン」, ニュートリ
  シー出版, 45-52, 2013. ISBN: 978-4-7813-0820-3
  内容:マイクロアレイ解析の話がメインです。実験デザインの重要性を述べて
  動遺伝子(DEG)検出法の組合せの重要性の話や、サンプル間クラスタリン
  ます。MAS5データを用いる場合は特に倍率変化で議論することも無意味で
  得られたマイクロアレイデータの場合には なぜ倍率変化でうまくいく傾向に

講習会、講義、講演資料

- **門田幸二**,「講義資料」, アグリバイオインフォマティクス教育研究プログラムの大学院講義科目:農学生命情報科学特論I, 東京大学(東京), 2014.07.02
  内容:教科書の3.3節と4.3節周辺。マッピングプログラムは大きくbowtieなどのbasic aligner (unspliced aligner)とtophatなどのsplice-aware aligner (spliced aligner)に大別されること。splice-aware alignerの基本的なイメージ。ゲノム配列既知の場合の遺伝子構造推定としてTophat-Cufflinksパイプラインの基本形を紹介。既知遺伝子(または転写物)の発現解析でよい場合は、トランスクリプトーム配列へのマッピングでよい。最近はSailfishやRNA-Skimなど、k-merに基づくalignment-freeな方法が注目されていることなど。研究目的別留意点として、遺伝子間比較の場合とサンプル間比較の場合、配列長補正、総リード数補正、RPKMなど。長い転写物ほどマップされるリード数が多い傾向をRで確認。GSE42212のヒトRNA-seqデータのFASTQファイル取得以降の一通りの解析。実際に行ったのは、カウントデータ取得以降のTCCパッケージを用いたサンプル間クラスタリング、発現変動遺伝子(DEG)同定。M-A plotのおさらい。結果の解釈。FDR、分布やモデルの説明。倍率変化でDEG同定を行う場合との比較。2コマ(2×90 min)分。

門田の本務である大学院講義(90分×18コマ=**27時間分**)スライドを含め、2013年秋以降のPDFファイルを簡単な解説つきで公開しています。

R中心ですがトランスクリプトーム
データ解析を一通り学びたい人は…

(Rで)塩基配列解析の情報をフル活用

科目名:農学生命情報科学特論I
内容:公共DB、チェックサム、QC、前処理、k-mer、アセンブリ、マッピング、RPKM、発現変動など。
実施日:2014.06.18、2014.06.25、2014.07.02

科目名:機能ゲノム学
内容:データ取得、正規化、クラスタリング、発現変動解析、多重比較問題、機能解析など。
実施日:2014.05.14、2014.05.21、2014.05.28、2014.06.04

科目名:ゲノム情報解析基礎
内容:Rの基礎。GC含量計算やCpG解析、上流配列解析、Rのバージョンの違いなど。
実施日:2014.04.09、2014.04.23、2014.04.30

これら3科目の講義資料を順番にみていくとよい

(Rで)塩基配列解析の情報をフル活用

(RobLoxBioC)の紹介および結果が変わらないことの確認までをやってもらった。2コマ(2×90 min)分。

- **門田幸二**,「講義資料」, アグリバイオインフォマティクス教育研究プログラムの大学院講義科目:ゲノム情報解析基礎, 東京大学(東京), 2014.04.30
  内容: Rで塩基配列解析を行うための基本的なところ。例題としてシロイヌナズナゲノムのCpG出現頻度を解析し考察。Rパッケージのインストール、エラーメッセージへの対処法、利用可能な関数の概観。sequence logosを主な講義内容とし、エントロピー計算や、なぜエントロピーをそのまま利用せずに情報量に変換するかの意義。subseq関数のオプションをうまく利用して効率的に目的のプロモーター配列領域を切り出して計算するやり方など。課題4はプログラムの一部を任意に変更する基礎的な能力を問うもの。他の例題の中に回答が存在するので、それを効率的に見つける能力を見ている。講義自体はスライド39までで、スライド40以降はうまくいかないこともあるという事例やRのバージョンの違いに気をつけろ的な話。「農学生命情報科学特論I」で改めて話す予定。1コマ(90 min)分。
- **門田幸二**,「講義資料」, アグリバイオインフォマティクス教育研究プログラムの大学院講義科目:ゲノム情報解析基礎, 東京大学(東京), 2014.04.23
  内容: Rで塩基配列解析を行うための基本的なところ。初心者が犯しがちなミス、プログラムの中身の説明、アノテーションファイルやmulti-FASTAファイルからの情報抽出、意図的にエラーを出させてエラーへの対処能力向上、GC含量計算やそのプログラム内部の説明、ヒトゲノムのCpG出現頻度を解析するための連続塩基出現頻度解析、BSgenomeパッケージとか。課題は、自分が解析したい入力ファイルの全体像を把握し、適切な列およびキーワードで効率よく情報収集するための練習問題レベルのものにしている。Rがいかに簡単であるかをわかってもらうことに重点を置いている。ただし、ヘッダー行でひっかけを作っており、目で見て明らかに回答がわかっている状況下でそれを正しく判断し適切なテンプレートプログラムを利用できるかを問うている。また、課題2では、ゲノム配列にもバージョンがあるということを認識してもらう。2コマ(2×90 min)分。
- **門田幸二**,「ウェブページと講義資料」, アグリバイオインフォマティクス教育研究プログラムの大学院講義科目:ゲノム情報解析基礎, 東京大学(東京), 2014.04.09
  内容: 初心者向けバイオインフォマティクス全般およびゲノム情報解析系のイントロダクションの話。Rのイントロダクションやこのウェブページの簡
- **門田幸二**,「比較トランスクリプトーム解析とその周辺
  よく分かる次世代シークエンサー解析ワークショップ
  内容: 初心者向けRNA-seqの話。主にカウントデータ

「ゲノム情報解析基礎」の講義資料はこちら

科目名:農学生命情報科学特論I
内容:公共DB、チェックサム、QC、前処理、k-mer、アセンブリ、マッピング、RPKM、発現変動など。
実施日:2014.06.18、2014.06.25、2014.07.02

科目名:機能ゲノム学
内容:データ取得、正規化、クラスタリング、発現変動解析、多重比較問題、機能解析など。
実施日:2014.05.14、2014.05.21、2014.05.28、2014.06.04

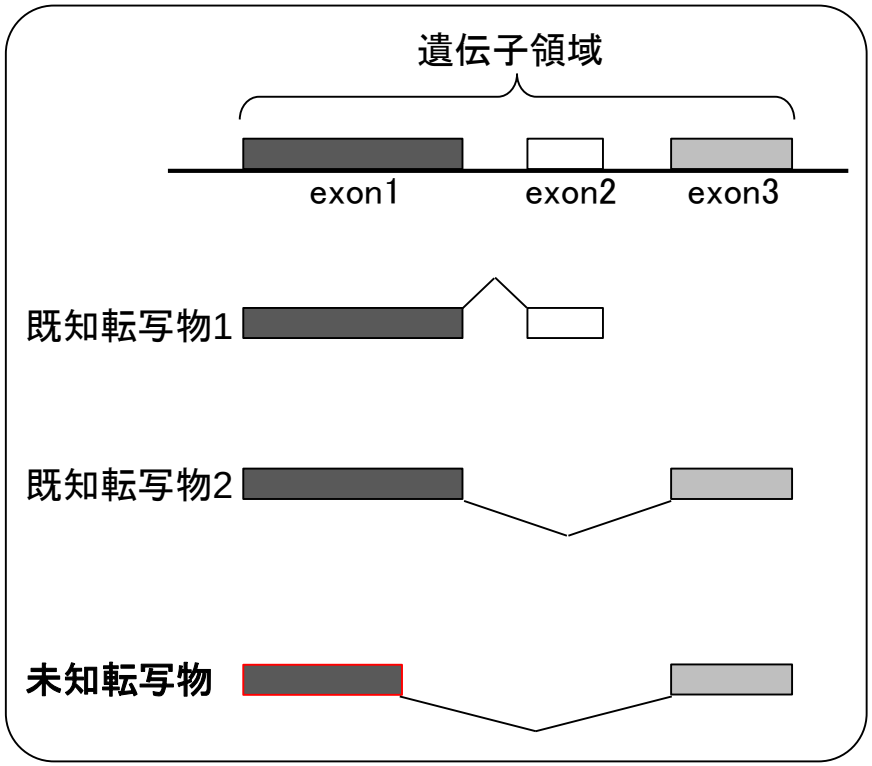
科目名:ゲノム情報解析基礎
内容:Rの基礎。GC含量計算やCpG解析、上流配列解析、Rのバージョンの違いなど。
実施日:2014.04.09、2014.04.23、2014.04.30

Contents

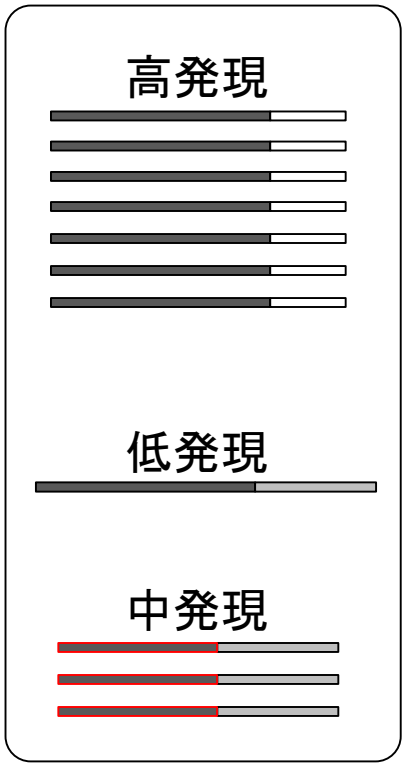
- 情報収集先
  - イルミナのウェビナー
  - (Rで)塩基配列解析
- **トランスクリプトーム解析技術の進展と展望**
  - Wet側
    - マイクロアレイ：3’発現アレイ → エクソンアレイ → トランスクリプトームアレイ
    - RNA-seq：Illumina short-readとPacBio long-read
  - Dry側
    - 遺伝子構造推定（ゲノム配列を利用）
    - データ解析手段（ウェブツール、Linux、R）
    - 転写物の発現量推定（トランスクリプトーム配列を利用）
    - 発現変動解析（Rパッケージを利用）

RNA-seq基本イメージのおさらい

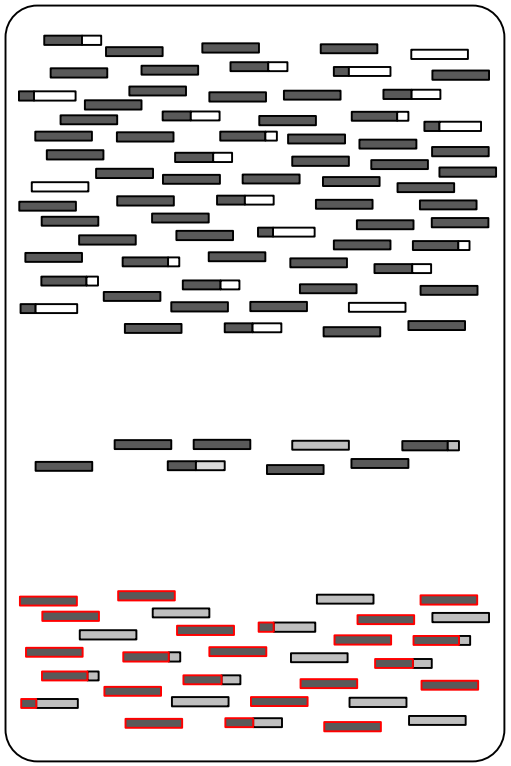
- 真の転写物情報：ある遺伝子領域中に既知転写物は2つ、**未知転写物**も1つ!
- 真の発現情報：既知転写物1（高発現）、既知転写物2（低発現）、未知転写物（中発現）



真の転写物情報



真の発現情報

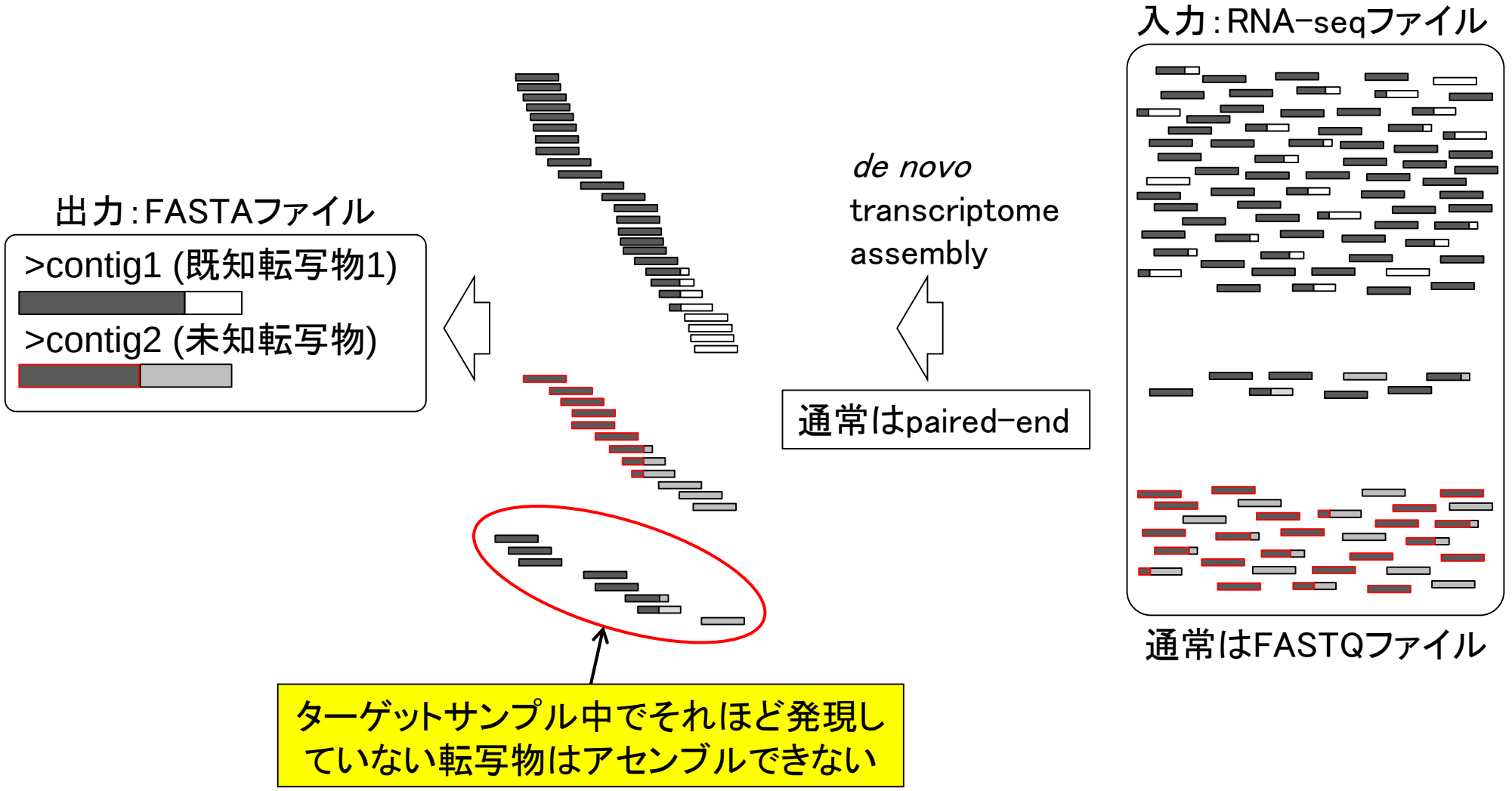


RNA-seqで得られるリード情報
（色は不明；single-endの場合）

どの転写物由来か分からない塩基配列情報のみがRNA-seqによって得られる。
これをもとに真の転写物情報や発現情報を得るのがRNA-seqデータ解析の目的

おさらい（RNA-seqの主な目的1）

- RNA-seqデータのみしか手元にない場合：トランスクリプトーム配列取得



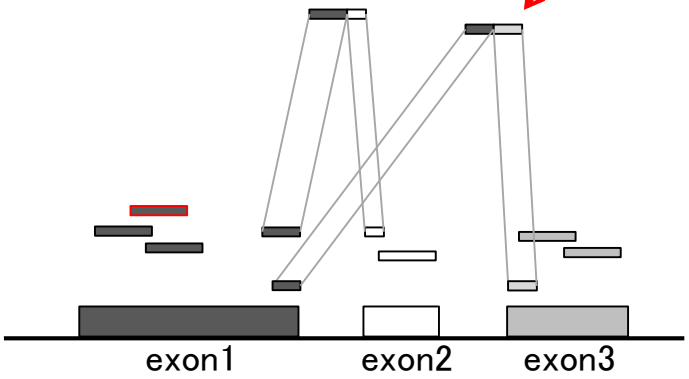
おさらい（RNA-seqの主な目的2）

- リファレンスとしてゲノム配列が利用可能な場合：新規転写物の同定

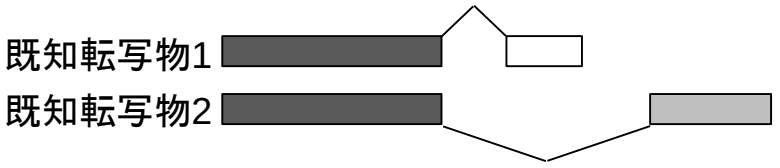
リファレンス配列を利用することで低発現転写物の遺伝子構造推定がde novo assemblyに比べて容易に

ジャンクションリードもマッピング可能

マッピング

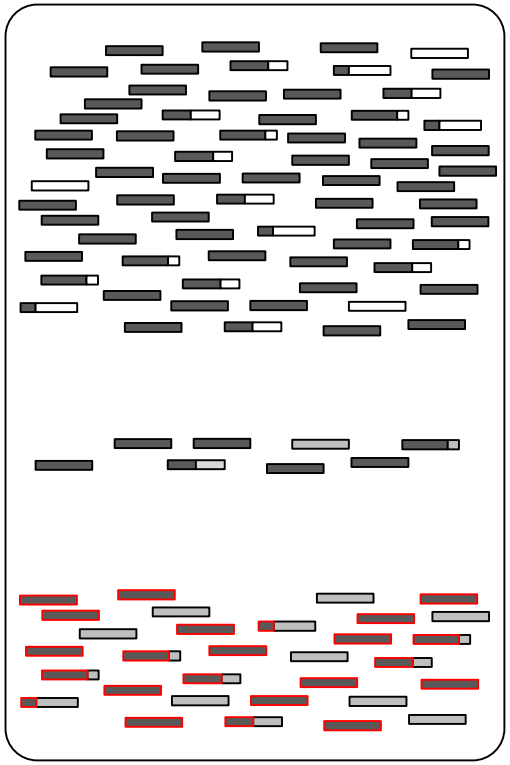


入力2：ゲノム配列



（入力3：アノテーション情報、既知遺伝子座標情報）

入力1：RNA-seqファイル



通常はFASTQファイル

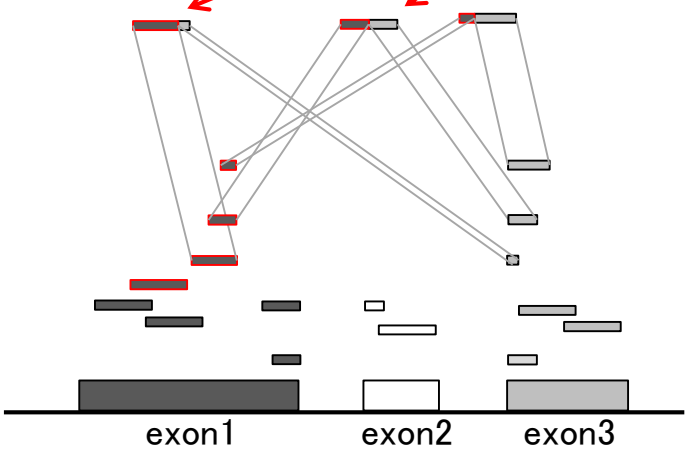
おさらい（RNA-seqの主な目的2）

- リファレンスとしてゲノム配列が利用可能な場合：新規転写物の同定

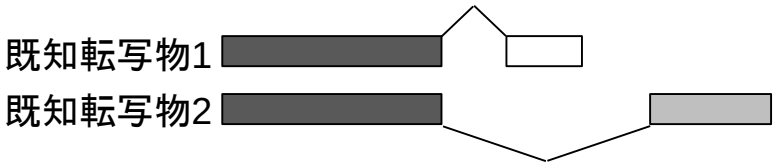
リファレンス配列を利用することで低発現転写物の遺伝子構造推定がde novo assemblyに比べて容易に

ジャンクションリードもマッピング可能

マッピング

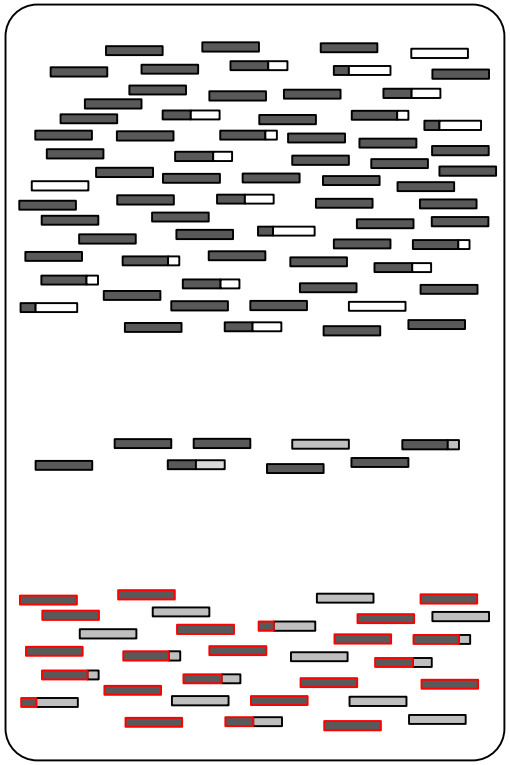


入力2：ゲノム配列



（入力3：アノテーション情報、既知遺伝子座標情報）

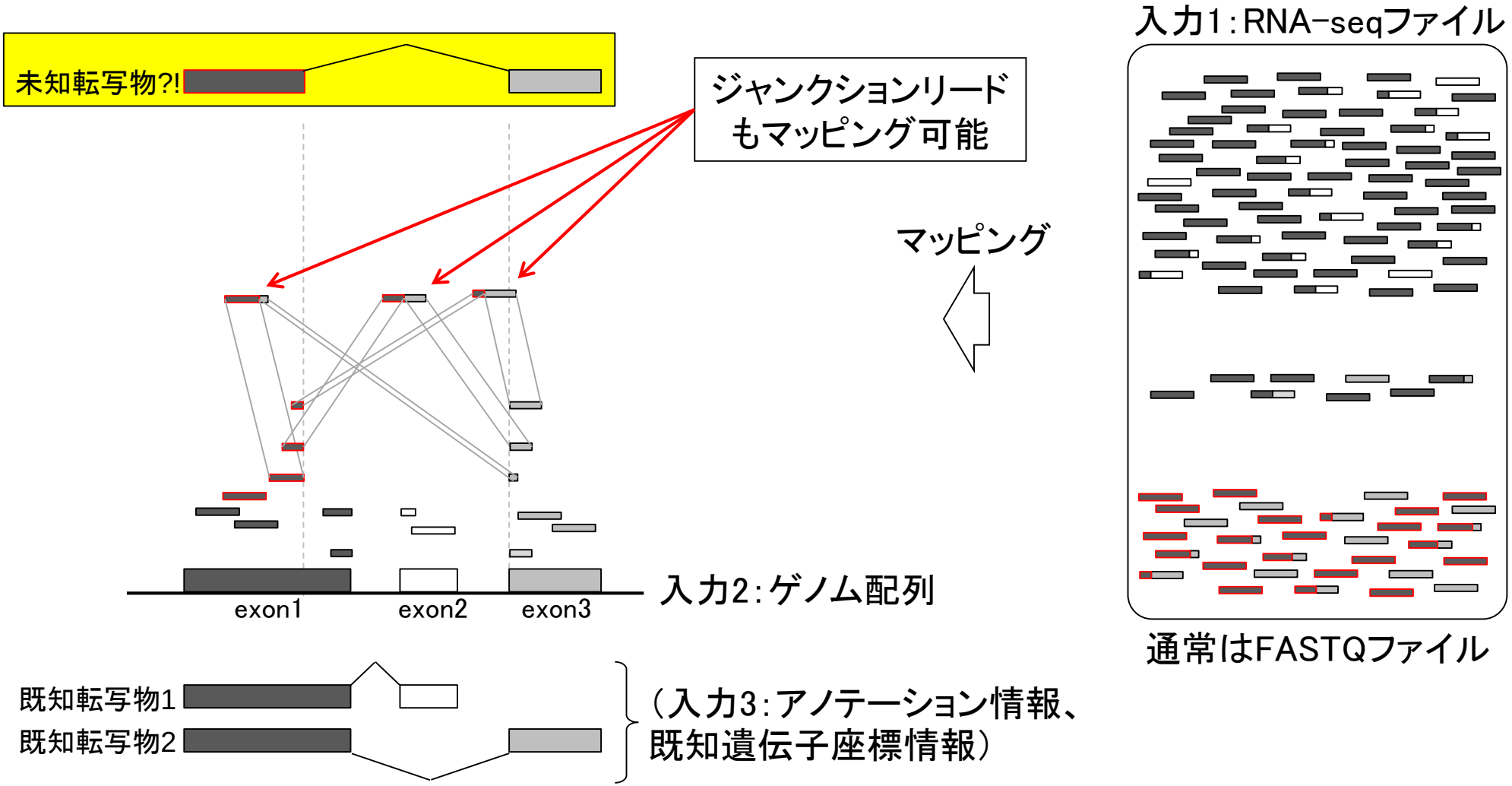
入力1：RNA-seqファイル



通常はFASTQファイル

おさらい（RNA-seqの主な目的2）

- リファレンスとしてゲノム配列が利用可能な場合：新規転写物の同定



Contents

- 情報収集先
  - イルミナのウェビナー
  - (Rで)塩基配列解析
- トランスクリプトーム解析技術の進展と展望
  - **Wet側**
    - **マイクロアレイ:3’発現アレイ → エクソンアレイ → トランスクリプトームアレイ**
    - **RNA-seq:Illumina short-readとPacBio long-read**
  - Dry側
    - 遺伝子構造推定(ゲノム配列を利用)
    - データ解析手段(ウェブツール、Linux、R)
    - 転写物の発現量推定(トランスクリプトーム配列を利用)
    - 発現変動解析(Rパッケージを利用)

トランスクリプトーム解析技術の進展(Wet)

- マイクロアレイの進展
  - 3’ 発現アレイ（2001年頃）
    - プローブが転写物の3’ 側に偏っていた（EST解析やcDNAプロジェクトの時代）
    - これまで蓄積された情報との比較が可能（未だに利用される所以）
  - エクソンアレイ（2007年頃）
    - モデル生物のエクソン領域を網羅的にカバー（例：GPL5188）
    - エクソンアレイデータを取り扱うバイオインフォマティシャンはそこそこいた
  - ヒトトランスクリプトームアレイ（2013年発売）
    - Long non-coding RNA (lncRNA)やイントロン領域のプローブも搭載
    - GSE54143 (Wang et al., *Science*, **344**: 310-313, 2014)

マイクロアレイも進展している。ただし解像度の高いトランスクリプトームアレイはヒト用のみ。モデル生物、非モデル生物を問わずに適用可能という点が多くの研究者がRNA-seqに移行している所以であろう。

トランスクリプトーム解析技術の進展(Wet)

- RNA-seqの進展
  - Illumina (short-read → medium-read)
    - MiSeq：15Gb | 25M (2,500万リード) | 300bp × 2
    - NextSeq 500：120Gb | 400M (4億リード) | 150bp × 2
    - HiSeq 2500：1,000Gb | 4B (40億リード) | 125bp × 2、Rapidモードで250bp × 2が可能?!
  - Pacific Biosciences (long-read)
    - PacBio RS II：リード数そこそこで10,000bp程度読める（詳細なスペックは不明）
    - ERP003225 (Sharon et al., Nat Biotechnol., 2013)：48万 circular-consensus (CCS)リード | 1,000bp
    - SRP036136 (Tilgner et al., PNAS, 2014)：70万CCSリード | 1,200bp

de novo transcriptome assembly周辺で四苦八苦する時代はそろそろ終焉。パーソナルトランスクリプトームの時代へ。

de novo transcriptome assembly

(Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル、バイオインフォマティクス～
(last modified 2014/07/07, since 2010)

What's new?

- このウェブページはフ
  すので、Rのインスト
- 2014年7月22日にイ
- 門田幸二 著シリーズ
- マップ後 | 配列長とカ
  ニックとして「floor(nr
  明したため「ceiling(n
- 2014年9月1日～12日
  東大農で開催します。
  TA申込みが全日程
  m(_ _)m。7/4の10:00
- 参考資料(講義、講習

- はじめに (last modifie
- 参考資料(講義、講習

- 前処理 | トリミング | アダプター配列除去(応用) | ShortRead(Morgan 2009)(last modified 2014/06/18)推奨 NEW
- 前処理 | トリミング | 指定した末端塩基数だけ除去(last modified 2013/06/15)
- アセンブル | について(last modified 2014/06/20) NEW
- アセンブル | ゲノム用(last modified 2014/07/0 NE
- アセンブル | トランスクリプトーム(転写物)用 mo
- マッピング | について(last modified 2014/07/0 NE
- マッピング | basic aligner(last modified 2014/06/24)
- マッピング | splice-aware aligner(last modified 2014/
- マッピング | Bisulfite sequencing用(last modified 201
- マッピング | (ESTレベルの長さの)contig(last modifie
- マッピング | 基礎(last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) |
- マッピング | single-end | ゲノム | basic aligner(応用) |
- マッピング | single-end | ゲノム | splice-aware aligner
- マップ後 | について(last modified 2013/06/19)
- マップ後 | 出力ファイル形式について(last modified 2
- マップ後 | 出力ファイルの読み込み | BAM形式(last
- マップ後 | 出力ファイルの読み込み | Bowtie形式(las
- マップ後 | 出力ファイルの読み込み | SOAP形式(last
- マップ後 | 出力ファイルの読み込み | htSeqTools (Pla
- マップ後 | カウント情報取得 | について(last modified
- マップ後 | カウント情報取得 | ゲノム | アノテーション
- マップ後 | カウント情報取得 | ゲノム | アノテーション
- マップ後 | カウント情報取得 | トランスクリプトーム | B
- マップ後 | 配列長とカウント数の関係(last modified 2

アセンブル | トランスクリプトーム(転写物)用 NEW

Rパッケージはおそらくありません。

プログラム:

プログラムやガイドラインもいくつかあり

- Multiple-k: Surget-Groba and Montoya-Burgos, Genome Res., 2010
- Trans-ABySS: Robertson et al., Nature Methods, 2010
- Rnnotator: Martin et al., BMC Genomics, 2010
- Trinity: Grabherr et al., Nat Biotechnol, 2011
- SOAPdenovo-trans: Luo et al., Gigascience, 2012
- Oases: Schulz et al., Bioinformatics, 2012
- EBARDenovo: Chu et al., Bioinformatics, 2013
- IDBA-tran: Peng et al., Bioinformatics, 2013

Review、ガイドライン、パイプライン系:

- Review: Martin and Wang, Nat Rev Genet., 2011
- ガイドライン: Haznedaroglu et al., BMC Bioinformatics, 2012
- Review: Góngora-Castillo, Nat Prod Rep., 2013
- ガイドライン: Yang and Smith, BMC Genomics, 2013
- ガイドライン: O'Neil et al., BMC Genomics, 2013
- ガイドライン: Feldmesser et al., BMC Genomics, 2014
- パイプライン: Melicher et al., BMC Genomics, 2014

Contents

- 情報収集先
  - イルミナのウェビナー
  - (Rで)塩基配列解析
- トランスクリプトーム解析技術の進展と展望
  - Wet側
    - マイクロアレイ：3’発現アレイ → エクソンアレイ → トランスクリプトームアレイ
    - RNA-seq：Illumina short-readとPacBio long-read
  - **Dry側**
    - **遺伝子構造推定（ゲノム配列を利用）**
    - データ解析手段（ウェブツール、Linux、R）
    - 転写物の発現量推定（トランスクリプトーム配列を利用）
    - 発現変動解析（Rパッケージを利用）

トランスクリプトーム解析技術の進展(Dry)

- 遺伝子構造推定(ゲノム配列を利用)
  - 概要
    - Paired-endリードのゲノムへのマッピング結果を入力として、転写される領域(遺伝子構造)を推定
    - 新規転写物(isoformsやsplice variants)の同定、転写物配列取得と同義
    - 多くのプログラムは、遺伝子レベルや転写物レベルの発現量まで出力
  - 手段
    - 複数サンプルデータの利用が基本
    - ドラフトゲノムでも適用可能なもの、アノテーションファイルなしで遺伝子構造推定を行うもの(de novo reconstruction)など多数存在
  - 評価基準
    - 既知転写物をどれだけ多く同定できるか？
    - 既知の偽遺伝子(pseudogenes)をどれだけレポートしないか？
    - 未知転写物の中にどれだけ偽陽性(false positives)が含まれないか？

代表的なTophat-Cufflinksパイプライン以外にもいろいろ試してみるとよいのでは…

トランスクリプトーム解析技術の進展(Dry)

- 解析 | 基礎 | 平均-分散プロット | Technical replicates(last modified 2014/02/18)
- 解析 | 基礎 | 平均-分散プロット | Biological replicates(last modified 2014/02/21)
- 解析 | 新規転写物同定(ゲノム配列を利用)(last modified 2014/07/08) NEW
- 解析 | 発現量推定(トランスクリプトーム配列を利用)(last modified 2014/07/08) NEW
- 解析 | クラスタリング | について(last modified 2014/02/05)
- 解析 | クラスタリング | サンプル間 | hclust(last modified 2014/06/30) NEW
- 解析 | クラスタリング | サ
- 解析 | クラスタリング | 遺
- 解析 | 発現変動 | ボアソ
- 解析 | 発現変動 | 負の二
- 解析 | 発現変動 | につい
- 解析 | 発現変動 | 2群間
- 解析 | 発現変動 | 2群間
- 解析 | 発現変動 | 2群間
- 解析 | 発現変動 | 2群間

解析 | 新規転写物同定(ゲノム配列を利用) NEW

reference-based methodsというカテゴリに含まれるものたちです。ゲノム配列にRNA-seqデータのマッピングを行ってどこに遺伝子領域があるかなどの座標(アノテーション)情報を取得する遺伝子構造推定用です。実質的にde novo transcriptome assemblyと目指すところは同じですが、やはりゲノムというリファレンス配列を用いるほうがより正確であるため、ゲノム配列が利用可能な場合は利用するのが一般的です。一般にこの種のプログラムは遺伝子構造推定だけでなく、発現量推定まで行ってくれます。とりあえずリストアップしただけのものもあったと思いますので、ゲノムにマップしないものがいくつか含まれているとは思います。

R用:

- Solas(Windows版はなさそう; 2010年以降アップデートなし): Richard et al., Nucleic Acids Res., 2010
- NSMAP(Windows版はなさそう;アノテーションファイルなしで実行): Xia et al., BMC Bioinformatics, 2011
- Sequgio: Suo et al., Bioinformatics, 2014
- spliceR(Cufflinksの出力をインプットにしているようだ): Vitting-Seerup et al., BMC Bioinformatics, 2014

R以外:

ざっと調べただけでも30種類以上あります

- Scripture: Guttman et al., Nat Biotechnol., 2010
- Cufflinks: Trapnell et al., Nat Biotechnol., 2010
- rQuant: Bohnert and Ratsch, Nucleic Acids Res., 2010
- STM: Surget-Groba and Montoya-Burgos, Genome Res., 2010
- ALEXA-seq: Griffith et al., Nat Methods, 2010
- MISO: Katz et al., Nat. Methods, 2010
- MMSEQ: Turro et al., Genome Biol., 2011
- IsoEM: Nicolae et al., Algorithms Mol. Biol., 2011

解析 | 新規転写物同定(ゲノム配列を利用) NEW

reference-based methodsというカテゴリに含まれるものたちです。ゲノム配列にRNA-seqデータのマッピングを行ってどこに遺伝子領域があるかなどの座標(アノテーション)情報を取得する遺伝子構造推定用です。実質的に de novo transcriptome assemblyと目指すところは同じですが、やはりゲノムというリファレンスがあった方がより正確であるため、ゲノム配列が利用可能な場合は利用するのが一般的です。一般に遺伝子構造推定だけでなく、発現量推定まで行ってくれます。とりあえずリストアップしたためますので、ゲノムにマップしないものがいくつか含まれているとは思います。

R用:

- Solas(Windows版はなさそう; 2010年以降アップデートなし): Richard et al., Nucl
- NSMAP(Windows版はなさそう;アノテーションファイルなしで実行): Xia et al., E
- Sequgio: Suo et al., Bioinformatics, 2014
- spliceR(Cufflinksの出力をインプットにしているようだ): Vitting-Seerup et al., BM

R以外:

- Scripture: Guttman et al., Nat Biotechnol., 2010
- Cufflinks: Trapnell et al., Nat Biotechnol., 2010
- rQuant: Bohnert and Ratsch, Nucleic Acids Res., 2010
- STM: Surget-Groba and Montoya-Burgos, Genome Res., 2010
- ALEXA-seq: Griffith et al., Nat Methods, 2010
- MISO: Katz et al., Nat. Methods, 2010

ここで示しているのは基本的にマッピングから遺伝子構造推定までの一連のパイプライン。内部的にどのマッピングプログラムを用いるかによって、さらに多数の組合せが原理的に可能。例えば、Tophat-Cufflinksパイプラインは、Tophatというジャンクションリードをマッピングできるプログラム出力結果を用いてCufflinksというアルゴリズムで遺伝子構造推定を行うものという理解でよい。

- Scripture: Guttman et al., Nat Biotechnol., 2010
- Cufflinks: Trapnell et al., Nat Biotechnol., 2010
- rQuant: Bohnert and Ratsch, Nucleic Acids Res., 2010
- STM: Surget-Groba and Montoya-Burgos, Genome Res., 2010
- ALEXA-seq: Griffith et al., Nat Methods, 2010
- MISO: Katz et al., Nat. Methods, 2010
- MMSEQ: Turro et al., Genome Biol., 2011
- IsoEM: Nicolae et al., Algorithms Mol. Biol., 2011
- IsoformEx: Kim et al., BMC Bioinformatics, 2011
- RSEM: Li and Dewey, BMC Bioinformatics, 2011
- SLIDE: Li et al., PNAS, 2011
- IQSeq: Du et al., PLoS One, 2012
- BitSeq: Glaus et al., Bioinformatics, 2012
- ARTADE2: Kawaguchi et al., Bioinformatics, 2012
- RD: Wan et al., Biostatistics, 2012
- Plntron: Pirola et al., BMC Bioinformatics, 2012
- CEM: Li and Jiang, Bioinformatics, 2012
- iReckon: Mezlini et al., Genome Res., 2013
- TrueSight: Li et al., Nucleic Acids Res., 2013(p3)
- PASTA: Tang et al., BMC Bioinformatics, 2013
- CLASS: Song and Florea, BMC Bioinformatics, 2013
- GeneScissors: Zhang et al., Bioinformatics, 2013
- PSGInfer: LeGault et al., Bioinformatics, 2013
- NURD: Ma et al., BMC Bioinformatics, 2013
- MITIE: Behr et al., Bioinformatics, 2013
- ORMAN: Dao et al., Bioinformatics, 2014
- UnSplicer: Burns et al., Nucleic Acids Res., 2014
- PennSeq: Hu et al., Nucleic Acids Res., 2014
- Parseq: Mirauta et al., Bioinformatics, 2014

トランスクリプトーム解析技術の進展(Dry)

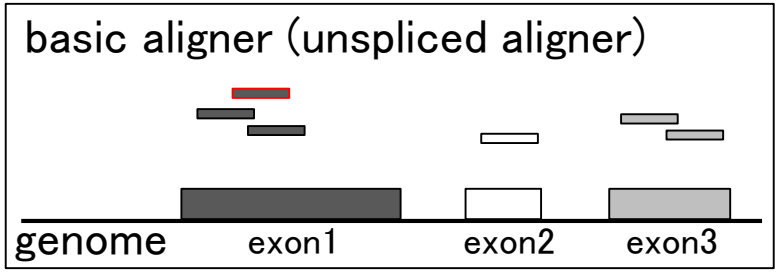
- アセンブル | トランスクリプトーム(転写物)用 (last modified 2014/07/08)
- マッピング | について (last modified 2014/07/08) NEW
- マッピング | basic aligner (last modified 2014/06/24) NEW
- マッピング | splice-aware aligner (last modified 2014/07/08) NEW
- マッピング | Bisulfite sequencing用 (last modified 2014/06/24) NEW
- マッピング | (ESTレベルの長さの)contig (last modified 2014/06/24) NE

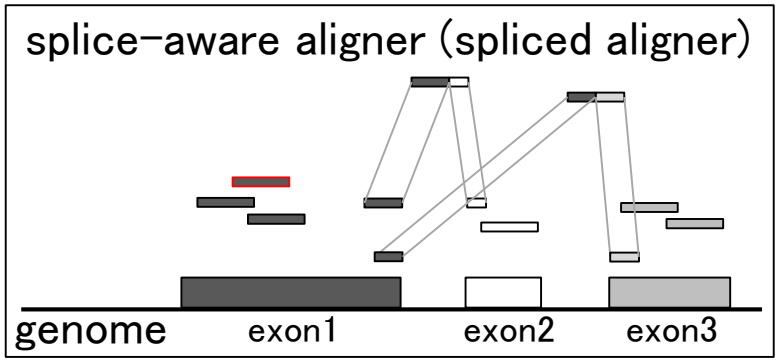
マッピング | splice-aware aligner NEW

basic alignerはジャンクションリード(junciton reads; spliced reads)のマッピングができません。splice-aware alignerは計算に時間がかかるものの、それらもマッピングしてくれます。SpliceMapは、リードの半分の長さをマップさせておいてそのアラインメントを拡張(extend)させる戦略を採用しています。いずれの方法も、バージョンアップがどんどんなされるプログラムは、アルゴリズム(計算手順)自体も変わっていたりしますので参考程度にしてください。大きな分類としては、seed-and-extend系とexon-first系に分けられるようです。exon-first系は、内部的にbasic-alignerを用いてざっくりとマップできるものをマップし、マップされなかったものがジャンクションリード候補としてリードを分割してマップされる場所を探すイメージです。seed-and-extend系は、前者に比べて計算時間がかかるものの、全リードを等価に取り扱うため、exon-first系にありがちな「本当はジャンクションリードなんだけどbasic alignerでのマッピング時にpseudogeneにマップされてしまう」ということはないようです(Garber et al., 2011)。その後multi-seed系の方法も提案されているようです(Gatto et al., 2014)。

プログラム:

- BLAT: Kent WJ, Genome Res., 2002 seed-and-extend系
- QPALMA: De Bona et al., Bioinformatics, 2008 seed-and-extend系
- TopHat: Trapnell et al., Bioinformatics, 2009 exon-first系
- RNA-MATE: Cloonan et al., Bioinformatics, 2009
- GSNAP: Wu et al., Bioinformatics, 2010 seed-and-extend系
- SpliceMap: Au et al., Nucleic Acids Res., 2010 seed-and-extend系
- Supersplat: Bryant et al., Bioinformatics, 2010
- MapSplice: Wang et al., Nucleic Acids Res., 2010 multi-seed系
- HMMSplicer: Dimon et al., PLoS One, 2010
- X-MATE: Wood et al., Bioinformatics, 2011
- RUM: Grant et al., Bioinformatics, 2011





例えばMapSplice実行結果をCufflinksの入力として与えるとMapSplice-Cufflinksパイプラインとなります

トランスクリプトーム解析技術の進展(Dry)

- アセンブル | トランスクリプトーム(転写物)用(last modified 2014/07/08)
- マッピング | について(last modified 2014/07/08) NEW
- マッピング | basic aligner(last modified 2014/06/24) NEW
- マッピング | splice-aware aligner(last modified 2014/07/08) NEW
- マッピング | Bisulfite sequencing用(last modified 2014/06/24) NEW
- マッピング | (ESTレベルの長さの)contig(last modified 2014/06/24) NE

マッピング | splice-aware aligner NEW

basic alignerはジャンクションリード(junciton reads; spliced reads)のマッピングができません。splice-aware alignerは計算に時間がかかるものの、それらもマッピングしてくれます。SpliceMapは、リードの半分の長さをマップさせておいてそのアラインメントを拡張(extend)させる戦略を採用しています。いずれの方法も、バージョンアップがどんどんなされるプログラムは、アルゴリズム(計算手順)自体も変わっていたりしますので参考程度にしてください。大きな分類としては、seed-and-extend系とexon-first系に分けられるようです。exon-first系は、内部的にbasic-alignerを用いてざっくりとマップできるものをマップし、マップされなかったものがジャンクションリード候補としてリードを分割してマップされる場所を探すイメージです。seed-and-extend系は、前者に比べて計算時間がかかるものの、全リードを等価に取り扱うため、exon-first系にありがちな「本当はジャンクションリードなんだけどbasic alignerでのマッピング時にpseudogeneにマップされてしまう」ということはないようです(Garber et al., 2011)。その後multi-seed系の方法も提案されているようです(Gatto et al., 2014)。

プログラム:

- BLAT: Kent WJ, Genome Res., 2002seed-and-extend系
- QPALMA: De Bona et al., Bioinformatics, 2008seed-and-extend系
- TopHat: Trapnell et al., Bioinformatics, 2009exon-first系
- RNA-MATE: Cloonan et al., Bioinformatics, 2009
- GSNAP: Wu et al., Bioinformatics, 2010seed-and-extend系
- SpliceMap: Au et al., Nucleic Acids Res., 2010seed-and-extend系
- Supersplat: Bryant et al., Bioinformatics, 2010
- MapSplice: Wang et al., Nucleic Acids Res., 2010multi-seed系
- HMMSplicer: Dimon et al., PLoS One, 2010
- X-MATE: Wood et al., Bioinformatics, 2011
- RUM: Grant et al., Bioinformatics, 2011

- BLAT: Kent WJ, Genome Res., 2002seed-and-extend系
- QPALMA: De Bona et al., Bioinformatics, 2008seed-and-extend系
- TopHat: Trapnell et al., Bioinformatics, 2009exon-first系
- RNA-MATE: Cloonan et al., Bioinformatics, 2009
- GSNAP: Wu et al., Bioinformatics, 2010seed-and-extend系
- SpliceMap: Au et al., Nucleic Acids Res., 2010seed-and-extend系
- Supersplat: Bryant et al., Bioinformatics, 2010
- MapSplice: Wang et al., Nucleic Acids Res., 2010multi-seed系
- HMMSplicer: Dimon et al., PLoS One, 2010
- X-MATE: Wood et al., Bioinformatics, 2011
- RUM: Grant et al., Bioinformatics, 2011
- RNASEQR: Chen et al., Nucleic Acids Res., 2012
- PASSion: Zhang et al., Bioinformatics, 2012
- SOAPsplice: Huang et al., Front Genet., 2011
- ContextMap: Bonfert et al., BMC Bioinformatics, 2012
- OSA: Hu et al., Bioinformatics, 2012
- STAR: Dobin et al., Bioinformatics, 2013
- TrueSight: Li et al., Nucleic Acids Res., 2013
- OLego: Wu et al., Nucleic Acids Res., 2013
- TopHat2: Kim et al., Genome Biol., 2013multi-seed系
- segemehl: Hoffmann et al., Genome Biol., 2014
- HSA: Bu et al., BMC Syst Biol., 2013
- FineSplice: Gatto et al., Nucleic Acids Res., 2014
- lack(segemehlの一部): Otto et al., Bioinformatics, 2014

FineSpliceは内部的にTopHat2を利用してsplice junction同定精度を向上させたもののようです

解析 | 新規転写物同定(ゲノム配列を利用) NE

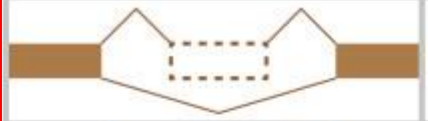
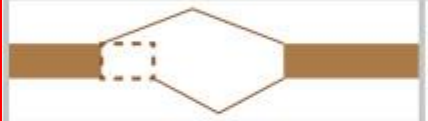
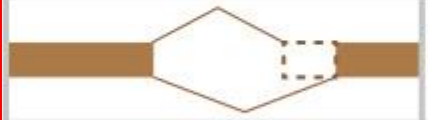
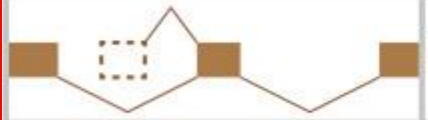
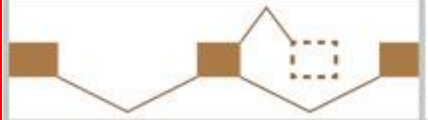
reference-based methodsというカテゴリに含まれるものたちです。ゲノム
てどこに遺伝子領域があるかなどの 座標(アノテーション)情報を取得す
novo transcriptome assemblyと目指すところは同じですが、 やはりゲノム
り正確であるため、ゲノム配列が利用可能な場合は利用するのが一般
伝子構造推定だけでなく、発現量推定まで行ってくれます。とりあえず
ますので、ゲノムにマップしないものがいくつか含まれているとは思いま

R用:

- Solas(Windows版はなさそう; 2010年以降アップデートなし): Ric
- NSMAP(Windows版はなさそう;アノテーションファイルなしで実
- Sequgio: Suo et al., Bioinformatics, 2014
- spliceR(Cufflinksの 出力をインプットにしているようだ): Vitting-

R以外:

- Scripture: Guttman et al., Nat Biotechnol., 2010
- Cufflinks: Trapnell et al., Nat Biotechnol., 2010
- rQuant: Bohnert and Ratsch, Nucleic Acids Res., 2010
- STM: Surget-Groba and Montoya-Burgos, Genome Res., 2010
- ALEXA-seq: Griffith et al., Nat Methods, 2010
- MISO: Katz et al., Nat. Methods, 2010
- MMSEQ: Turro et al., Genome Biol., 2011
- IsoEM: Nicolae et al., Algorithms Mol. Biol., 2011

	Event type	No. of events
	Exon skipping/inclusion (ESI)	1,619
	Mult. exon skipping/inclusion (MESI)	190
	Intron skipping/inclusion ISI	256
	Alternative 5' splice site A5	755
	Alternative 3' splice site A3	733
	Alternative transcription start site ATSS	1,381
	Alternative transcription termination site ATTS	1,125
	Mutually exclusive exons MEE	22
	All events	6,121

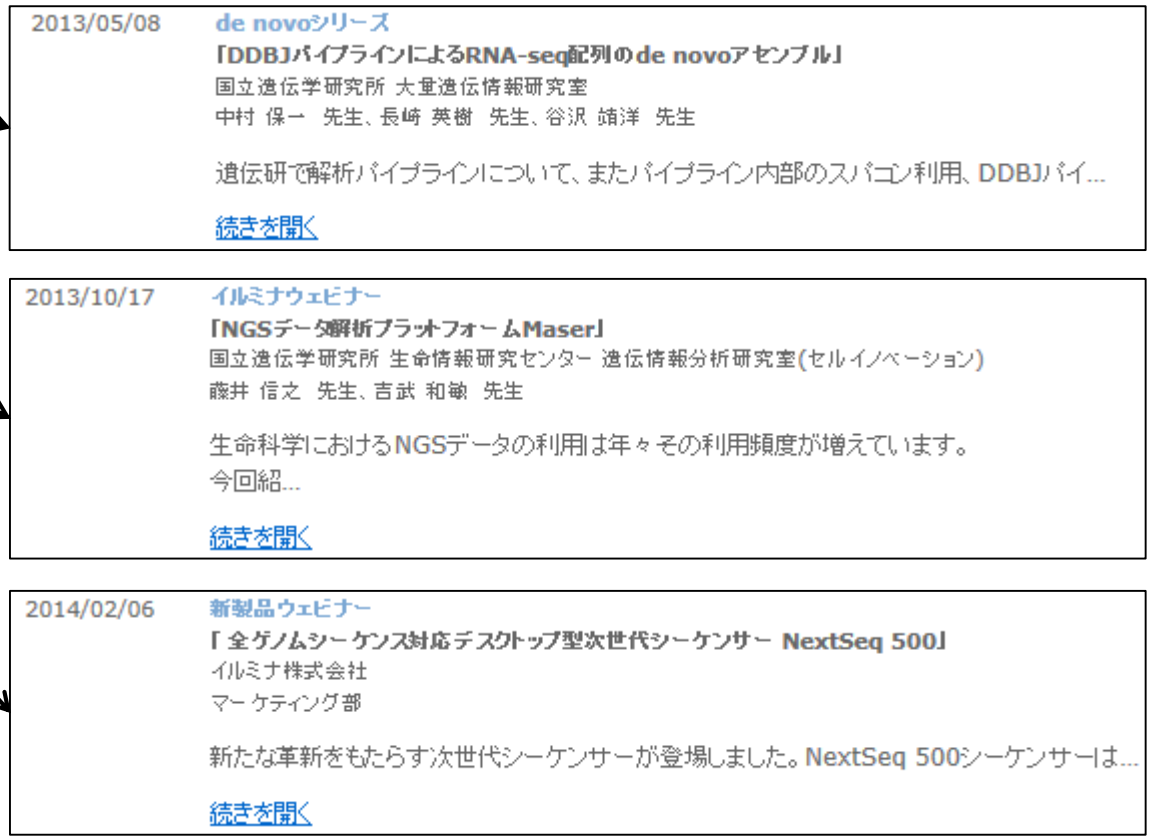
Cufflinksなどの遺伝子構造推定結果を入力として、Alternative Splicingの分類を行うRパッケージもあります

Contents

- 情報収集先
  - イルミナのウェビナー
  - (Rで)塩基配列解析
- トランスクリプトーム解析技術の進展と展望
  - Wet側
    - マイクロアレイ：3’発現アレイ → エクソンアレイ → トランスクリプトームアレイ
    - RNA-seq：Illumina short-readとPacBio long-read
  - Dry側
    - 遺伝子構造推定（ゲノム配列を利用）
    - **データ解析手段（ウェブツール、Linux、R）**
    - 転写物の発現量推定（トランスクリプトーム配列を利用）
    - 発現変動解析（Rパッケージを利用）

トランスクリプトーム解析技術の進展(Dry)

- データ解析手段(ウェブツール、Linux、R)
  - DDBJパイプライン
  - Maser
  - BaseSpace
  - …



2013/05/08 de novoシリーズ
「DDBJパイプラインによるRNA-seq配列のde novoアセンブル」
国立遺伝学研究所 大量遺伝情報研究室
中村 保一 先生、長崎 英樹 先生、谷沢 靖洋 先生

遺伝研で解析パイプラインについて、またパイプライン内部のスパコン利用、DDBJパイ...

続きを開く

2013/10/17 イルミナウェビナー
「NGSデータ解析プラットフォームMaser」
国立遺伝学研究所 生命情報研究センター 遺伝情報分析研究室(セルイノベーション)
藤井 信之 先生、吉武 和敏 先生

生命科学におけるNGSデータの利用は年々その利用頻度が増えています。
今回紹...

続きを開く

2014/02/06 新製品ウェビナー
「全ゲノムシーケンス対応デスクトップ型次世代シーケンサー NextSeq 500」
イルミナ株式会社
マーケティング部

新たな革新をもたらす次世代シーケンサーが登場しました。NextSeq 500シーケンサーは...

続きを開く

これらのウェブツールを利用することで、Tophat-Cufflinksなどのメジャーなパイプラインを Linux-free (Linuxを用いずに、という意味)で利用可能。

解析 | 新規転写物同定(ゲノム配列を利用) NEW

reference-based methodsというカテゴリに含まれるものたちです。ゲノム配列にRNA-seqデータのマッピングを行ってどこに遺伝子領域があるかなどの座標(アノテーション)情報を取得する遺伝子構造推定用です。実質的に1
novo transcriptome assemblyと目指すところは同じですが、やはりゲノムというリファレンり正確であるため、ゲノム配列が利用可能な場合は利用するのが一般的です。一般に
伝子構造推定だけでなく、発現量推定まで行ってくれます。とりあえずリストアップしたますので、ゲノムにマップしないものがいくつか含まれているとは思います。

R用:

- Solas(Windows版はなさそう; 2010年以降アップデートなし): Richard et al., Nucl
- NSMAP(Windows版はなさそう;アノテーションファイルなしで実行): Xia et al., I
- Sequgio: Suo et al., Bioinformatics, 2014
- spliceR(Cufflinksの出力をインプットにしているようだ): Vitting-Seerup et al., BM

R以外:

- Scripture: Guttman et al., Nat Biotechnol., 2010
- Cufflinks: Trapnell et al., Nat Biotechnol., 2010
- rQuant: Bohnert and Ratsch, Nucleic Acids Res., 2010
- STM: Surget-Groba and Montoya-Burgos, Genome Res., 2010
- ALEXA-seq: Griffith et al., Nat Methods, 2010
- MISO: Katz et al., Nat. Methods, 2010
- MMSEQ: Turro et al., Genome Biol., 2011
- IsoEM: Nicolae et al., Algorithms Mol. Biol., 2011

- Scripture: Guttman et al., Nat Biotechnol., 2010
- Cufflinks: Trapnell et al., Nat Biotechnol., 2010
- rQuant: Bohnert and Ratsch, Nucleic Acids Res., 2010
- STM: Surget-Groba and Montoya-Burgos, Genome Res., 2010
- ALEXA-seq: Griffith et al., Nat Methods, 2010
- MISO: Katz et al., Nat. Methods, 2010
- MMSEQ: Turro et al., Genome Biol., 2011
- IsoEM: Nicolae et al., Algorithms Mol. Biol., 2011

- IsoformEx: Kim et al., BMC Bioinformatics, 2011
- RSEM: Li and Dewey, BMC Bioinformatics, 2011
- SLIDE: Li et al., PNAS, 2011
- IQSeq: Du et al., PLoS One, 2012
- BitSeq: Glaus et al., Bioinformatics, 2012
- ARTADE2: Kawaguchi et al., Bioinformatics, 2012
- RD: Wan et al., Biostatistics, 2012
- Plntron: Pirola et al., BMC Bioinformatics, 2012
- CEM: Li and Jiang, Bioinformatics, 2012
- iReckon: Mezlini et al., Genome Res., 2013
- TrueSight: Li et al., Nucleic Acids Res., 2013(p3)
- PASTA: Tang et al., BMC Bioinformatics, 2013
- CLASS: Song and Florea, BMC Bioinformatics, 2013
- GeneScissors: Zhang et al., Bioinformatics, 2013
- PSGInfer: LeGault et al., Bioinformatics, 2013
- NURD: Ma et al., BMC Bioinformatics, 2013
- MITIE: Behr et al., Bioinformatics, 2013
- ORMAN: Dao et al., Bioinformatics, 2014
- UnSplicer: Burns et al., Nucleic Acids Res., 2014
- PennSeq: Hu et al., Nucleic Acids Res., 2014
- Parseq: Mirauta et al., Bioinformatics, 2014

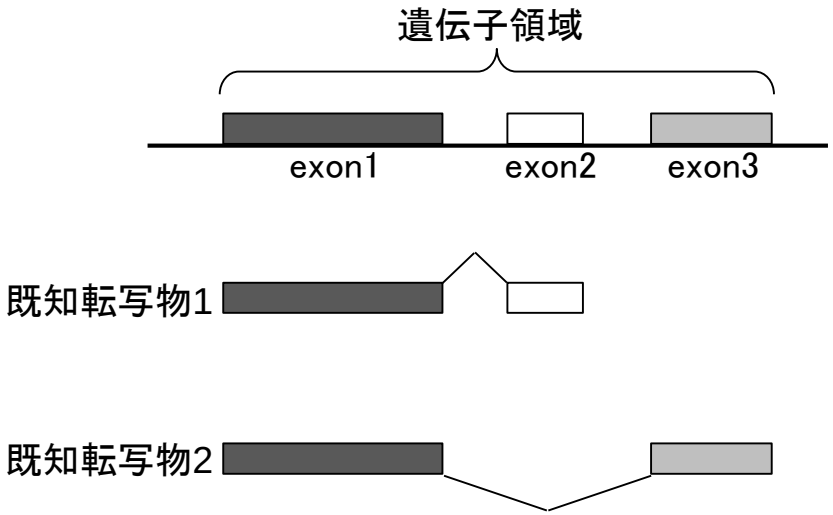
Cufflinks (Trapnell et al., 2010)以外にも多くのプログラムが開発されており、それらのほとんどがLinux上で動作。
→ Linuxをマスターする意義

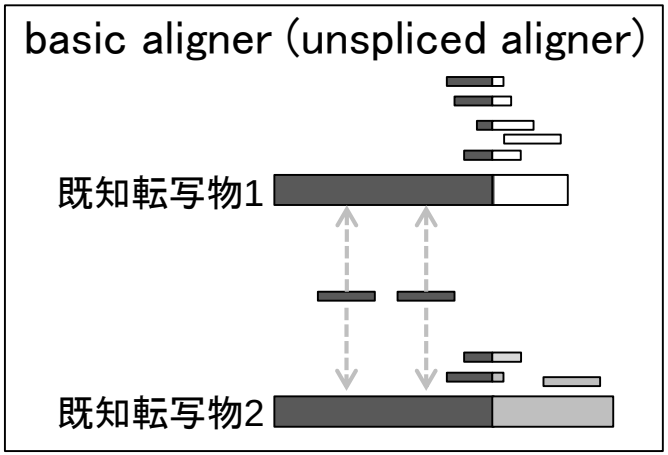
Contents

- 情報収集先
  - イルミナのウェビナー
  - (Rで)塩基配列解析
- トランスクリプトーム解析技術の進展と展望
  - Wet側
    - マイクロアレイ：3’発現アレイ → エクソンアレイ → トランスクリプトームアレイ
    - RNA-seq：Illumina short-readとPacBio long-read
  - Dry側
    - 遺伝子構造推定（ゲノム配列を利用）
    - データ解析手段（ウェブツール、Linux、R）
    - **転写物の発現量推定（トランスクリプトーム配列を利用）**
    - 発現変動解析（Rパッケージを利用）

トランスクリプトーム解析技術の進展(Dry)

- 転写物の発現量推定(トランスクリプトーム配列を利用)
  - 概要(2013年の論文までのトレンド)
    - ゲノム配列ではなくトランスクリプトーム配列にマップし、転写物ごとの発現量を推定
  - 手段
    - Tophat2などのsplice-aware alignerではなく、bowtie2などのbasic alignerを用いる





リファレンスがトランスクリプトーム配列なのでジャンクションリードかどうかは無関係。ただし、shared exon由来リードが複数転写物にマップされるので、その情報をできるだけ保持する必要がある。

トランスクリプトーム解析技術の進展(Dry)

- 解析 | 基礎 | 平均-分散プロット | Technical replicates(last modified 2014/02/18)
- 解析 | 基礎 | 平均-分散プロット | Biological replicates(last modified 2014/02/21)
- 解析 | 新規転写物同定(ゲノム配列を利用)(last modified 2014/07/08) NEW
- 解析 | 発現量推定(トランスクリプトーム配列を利用)(last modified 2014/07/08) NEW
- 解析 | クラスタリング | について(last modified 2014/02/25)
- 解析 | クラスタリング | サンプル間 | hclust(last modified 2014/06/30) NEW
- 解析 |
- 解析 |
- 解析 |
- 解析 |
- 解析 |
- 解析 |
- 解析 |
- 解析 |
- 解析 |

解析 | 発現量推定(トランスクリプトーム配列を利用) NEW

新規転写物(新規isoform)の発見などが目的でなく、既知転写物の発現量を知りたいだけの場合には、やたらと時間がかかるゲノム配列へのマッピングを避けるのが一般的です。有名なCufflinksも一応GTF形式のアノテーションファイルを与えることでゲノム全体にマップするのを避けるモードがあるらしいので、一応リストアップしています。転写物へのマッピングの場合には、splice-aware alignerを用いたジャンクションリードのマッピングを行う必要がないので、高速にマッピング可能なbasic alignerで十分です。但し、複数個所にマップされるリードは考慮する必要があり、確率モデルのパラメータを最尤法に基づいて推定するexpectation-maximization (EM)アルゴリズムがよく用いられます。マッピングを行わずに、k-merを用いてalignment-freeで行う発現量推定を行うSailfishやRNA-Skimは従来法に比べて劇的に高速化がなされているようです。間違いがいくつか含まれているとは思います。2014年6月に調べた結果をリストアップします:

プログラム:

- Cufflinks: Trapnell et al., Nat Biotechnol., 2010
- NEUMA: Lee et al., Nucleic Acids Res., 2011
- IsoEM: Nicolae et al., Algorithms Mol. Biol., 2011
- RSEM: Li and Dewey, BMC Bioinformatics, 2011
- eXpress: Roberts and Pachter, Nat Methods, 2013
- ReXpress: Roberts et al., Bioinformatics, 2013
- TIGAR: Nariai et al., Bioinformatics, 2013
- eXpress-D: Roberts et al., BMC Bioinformatics, 2013
- Sailfish: Patro et al., Nat Biotechnol., 2014
- RNA-Skim: Zhang and Wang, Bioinformatics, 2014

転写物配列(またはアノテーション)情報は日々更新されている。新規isoformが同定されるたびにマッピングをやり直さなくてもいいように、差分のみを取り扱うことでisoformレベルの発現量再推定を実行可能な時代に。

トランスクリプトーム解析技術の進展(Dry)

- 解析 | 基礎 | 平均-分散プロット | Technical replicates(last modified 2014/02/18)
- 解析 | 基礎 | 平均-分散プロット | Biological replicates(last modified 2014/02/21)
- 解析 | 新規転写物同定(ゲノム配列を利用)(last modified 2014/07/08) NEW
- 解析 | 発現量推定(トランスクリプトーム配列を利用)(last modified 2014/07/08) NEW
- 解析 | クラスタリング | について(last modified 2014/02/05)
- 解析 | クラスタリング | サンプル間 | bclust(last modified 2014/06/30) NEW
- 解析 |
- 解析 |
- 解析 |
- 解析 |
- 解析 |
- 解析 |
- 解析 |
- 解析 |
- 解析 |

解析 | 発現量推定(トランスクリプトーム配列を利用) NEW

新規転写物(新規isoform)の発見などが目的でなく、既知転写物の発現量を知りたいだけの場合には、やたらと時間がかかるゲノム配列へのマッピングを避けるのが一般的です。有名なCufflinksも一応GTF形式のアノテーションファイルを与えることでゲノム全体にマップするのを避けるモードがあるらしいので、一応リストアップしています。転写物へのマッピングの場合には、splice-aware alignerを用いたジャンクションリードのマッピングを行う必要がないので、高速にマッピング可能なbasic alignerで十分です。但し、複数個所にマップされるリードは考慮する必要があり、確率モデルのパラメータを最尤法に基づいて推定するexpectation-maximization (EM)アルゴリズムがよく用いられます。マッピングを行わずに、k-merを用いてalignment-freeで行う発現量推定を行うSailfishやRNA-Skimは従来法に比べて劇的に高速化がなされているようです。間違いがいくつか含まれているとは思います。2014年6月に調べた結果をリストアップします:

プログラム:

- Cufflinks: Trapnell et al., Nat Biotechnol., 2010
- NEUMA: Lee et al., Nucleic Acids Res., 2011
- IsoEM: Nicolae et al., Algorithms Mol. Biol., 2011
- RSEM: Li and Dewey, BMC Bioinformatics, 2011
- eXpress: Roberts and Pachter, Nat Methods, 2013
- ReXpress: Roberts et al., Bioinformatics, 2013
- TIGAR: Nariai et al., Bioinformatics, 2013
- eXpress-D: Roberts et al., BMC Bioinformatics, 2013
- Sailfish: Patro et al., Nat Biotechnol., 2014
- RNA-Skim: Zhang and Wang, Bioinformatics, 2014

トランスクリプトーム配列へのマッピングはゲノムに比べて早いことは間違いないが、マッピング自体が律速であることは変わりない。最近は*k*-merに基づくalignment-freeな定量法が注目されはじめている。

トランスクリプトーム解析技術の進展(Dry)

Bioinformatics. 2014 Jun 15;30(12):i283-i292. doi: 10.1093/bioinformatics/btu288.

RNA-Skim: a rapid method for RNA-Seq quantification at transcript level.

Zhang Z, Wang W.

⊕ Author information

Abstract

MOTIVATION: RNA-Seq technique has been demonstrated as a revolutionary means for exploring transcriptome because it provides deep coverage and base pair-level resolution. RNA-Seq quantification is proven to be an efficient alternative to Microarray technique in gene expression study, and it is a critical component in RNA-Seq differential expression analysis. Most existing RNA-Seq quantification tools require the alignments of fragments to either a genome or a transcriptome, entailing a time-consuming and intricate alignment step. To improve the performance of RNA-Seq quantification, an alignment-free method, Sailfish, has been recently proposed to quantify transcript abundances using all k-mers in the transcriptome, demonstrating the feasibility of designing an efficient alignment-free method for transcriptome quantification. Even though Sailfish is substantially faster than alternative alignment-dependent methods such as Cufflinks, using all k-mers in the transcriptome quantification impedes the scalability of the method.

RESULTS: We propose a novel RNA-Seq quantification method, RNA-Skim, which partitions the transcriptome into disjoint transcript clusters based on sequence similarity, and introduces the notion of sig-mers, which are a special type of k-mers uniquely associated with each cluster. We demonstrate that the sig-mer counts within a cluster are sufficient for estimating transcript abundances with accuracy comparable with any state-of-the-art method. This enables RNA-Skim to perform transcript quantification on each cluster independently, reducing a complex optimization problem into smaller optimization tasks that can be run in parallel. As a result, RNA-Skim uses <4% of the k-mers and <10% of the CPU time required by Sailfish. It is able to finish transcriptome quantification in <10 min per sample by using just a single thread on a commodity computer, which represents >100 speedup over the state-of-the-art alignment-based methods, while delivering comparable or higher accuracy. Availability and implementation: The software is available at http://www.csbio.unc.edu/rs.

CONTACT: weiwang@cs.ucla.edu Supplementary information: Supplementary data are available at Bioinformatics online.

これまでほぼ丸一日かかっていた計算が数分で完了する時代になりつつある。
1 day = 60*60*24 = 86,400 seconds

Zhang and Wang, *Bioinformatics*, 2014のTable 3

参考

トランスクリプトーム解析技術の進展(Dry)

- RNA-Skim (Zhang and Wang, *Bioinformatics*, 2014)
  - 概要
    - Preparationステージ：転写物集合を特定の*k*-mer (*sig-mer*)からなる小さなクラスターに分割。Sailfishは全ての*k*-merを利用するのに対して、RNA-Skimではそのクラスター中にのみ存在する*k*-merをsig-merとして利用。
    - Quantificationステージ：RNA-seq中のsig-merをカウント。(複数個所にマップされるリードに相当する)クラスター中の複数転写物で共有されるsig-merはもちろん存在するが、それらの割り振り問題はクラスター内で完結。

Preparationステージは予め計算しておけるため、新たなRNA-seqサンプルの定量化も高速に実行可能。ただし、現状ではReExpress (Roberts et al., 2013)のようなリファレンス配列のアップデートに対応する仕様にはなっていない。おそらく今後、新規isoformなどが得られた場合に差分のみPreparationステージで再計算するような改良版が出てくるのであろう。

k-merとは、*k*塩基からなる部分配列のこと。ヒトゲノム中にCpGという2連続塩基が期待値よりも非常に低いことを調べる2連続塩基の出現頻度解析は*k*=2の場合の*k*-merを利用した解析に相当。詳細は、20140423および20140625の講義資料を参照のこと。

Contents

- 情報収集先
  - イルミナのウェビナー
  - (Rで)塩基配列解析
- トランスクリプトーム解析技術の進展と展望
  - Wet側
    - マイクロアレイ：3’発現アレイ → エクソンアレイ → トランスクリプトームアレイ
    - RNA-seq：Illumina short-readとPacBio long-read
  - Dry側
    - 遺伝子構造推定（ゲノム配列を利用）
    - データ解析手段（ウェブツール、Linux、R）
    - 転写物の発現量推定（トランスクリプトーム配列を利用）
    - **発現変動解析（Rパッケージを利用）**

トランスクリプトーム解析技術の進展(Dry)

- 発現変動解析（Rパッケージを利用）
  - 入力：カウントデータ
    - **遺伝子**発現行列のような数値行列
    - 整数値からなる**遺伝子**領域上にマップされたリード数
  - 出力：発現変動遺伝子リスト（*p*-valueや*q*-value）やM-A plot

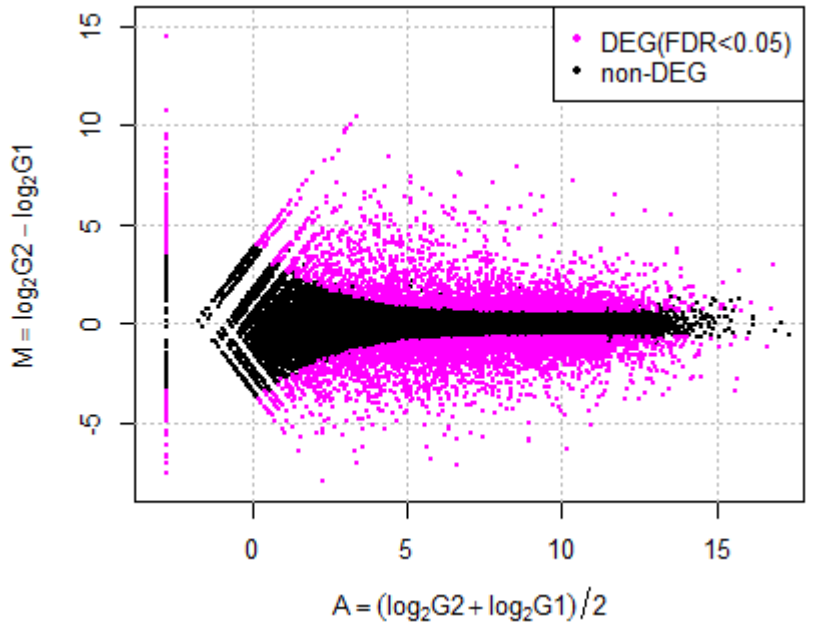
入力：カウントデータ

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

G1群　G2群

発現変動解析

出力：M-A plot



比較するグループ間(例：G1群 対 G2群)で発現に違いのある**遺伝子**(Differentially Expressed **Genes**; DEGs)の検定結果をレポート

トランスクリプトーム解析技術の進展(Dry)

トランスクリプトーム解析技術の進展(Dry)

- TCC (Sun et al., *BMC Bioinformatics*, 2013)
  - **TCCは内部的に既存パッケージ(edgeR, DESeq, and baySeq)中の関数を利用。既存パッケージ中のオリジナルの手順を繰り返し実行することで、データ正規化精度向上を実現。オリジナルの手順のみの場合に比べてより感度・特異度の高いDEG検出結果を得ることができる。**
  - TCC原著論文中では、edgeR, DESeq, baySeqパッケージ中の関数を自在に組み合わせて実行し、2群間比較の場合のみで性能評価している。推奨は以下の通り:
    - Biological replicatesありの場合:edgeR中の関数のみからなるiDEGES/edgeR正規化法
    - Biological replicatesなしの場合:DESeq中の関数のみからなるiDEGES/DESeq正規化法

  - **実質的には、より頑健なiterative edgeRやiterative DESeqを簡単に実行できるパッケージがTCC**という理解で差支えない。
  - 2013年7月の論文publish以降も継続的にアップデートしています
    - 多群間比較やpaired dataへの対応など、解析可能な実験デザインを拡張
    - DESeq2対応もほぼ完了
    - サンプル間クラスタリング用関数やマイクロアレイデータ用組織特異的発現パターン検出法ROKUの実装
    - ドキュメントが充実(TCC ver. 1.4.0で74ページに!)

compcodeRによる性能評価でもTCCの優位性を確認済

- 解析 | 発現変動 | について(last modified 2014/07/10) NEW
- 解析 | 発現変動 | 2群間 | 対応なし | について(last modified 2014/07/10) NEW
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun 2013)(last modified

解析 | 発現変動 | 2群間 | 対応なし | について NEW

実験デ

Aさん
Bさん
Cさん
Dさん
Eさん
Fさん
Gさん

2014年

R用:

R用:

- DEGSeq: Wang e
- edgeR: Robinson
- GPseq: Srivastava
- baySeq: Hardcast
- DESeq: Anders a
- DESeq2: Anders
- NBPSeq: Di et al.
- BBSeq: Zhou et a
- NOISeq: Tarazon
- PoissonSeq: Li et
- SAMseq: Li and
- easyRNASeq: De
- DSGseq: Wang et
- sSeq: Yu et al., B
- TCC: Sun et al., E
- tweeDEseq: Esna
- NPEBseq: Bi et a
- DER Finder: Fraz
- Characteristic Dir
- edgeR-robust: Zh
- ShrinkBayes: Var

http://bioconductor.org/packages/release/bioc/html/TCC.html

TCC

TCC: Differential expression analysis for tag count data with robust normalization strategies

Bioconductor version: Release (2.14)

This package provides a series of functions for performing differential expression analysis from RNA-seq count data using robust normalization strategy (called DEGES). The basic idea of DEGES is that potential differentially expressed genes or transcripts (DEGs) among compared samples should be removed before data normalization to obtain a well-ranked gene list where true DEGs are top-ranked and non-DEGs are bottom ranked. This can be done by performing a multi-step normalization strategy (called DEGES for DEG elimination strategy). A major characteristic of TCC is to provide the robust normalization methods for several kinds of count data (two-group with or without replicates, multi-group/multi-factor, and so on) by virtue of the use of combinations of functions in depended packages.

Author: Jianqiang Sun, Tomoaki Nishiyama, Kentaro Shimizu, and Koji Kadota

Maintainer: Jianqiang Sun <wukong at bi.a.u-tokyo.ac.jp>, Tomoaki Nishiyama <tomoakin at staff.kanazawa-u.ac.jp>

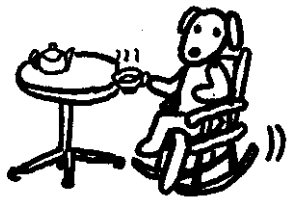
...

Documentation

PDF	R Script	TCC
PDF		Reference Manual
Text		NEWS

まとめ

- 情報収集先
  - イルミナのウェビナー
  - (Rで)塩基配列解析
- トランスクリプトーム解析技術の進展と展望
  - Wet側
    - マイクロアレイ:3’発現アレイ → エクソンアレイ → トランスクリプトームアレイ
    - RNA-seq:Illumina short-readとPacBio long-read
  - Dry側
    - 遺伝子構造推定(ゲノム配列を利用)
    - データ解析手段(ウェブツール、Linux、R)
    - 転写物の発現量推定(トランスクリプトーム配列を利用)
    - 発現変動解析(Rパッケージを利用)



トランスクリプトーム解析分野の最新状況を概観(したつもり)。高解像度化、高速化、高精度化、簡便化、…。進展が早すぎて半年後には違うことを言っているかもしれません。あしからずm(_ _)m

謝辞

共同研究者

清水 謙多郎　先生（東京大学・大学院農学生命科学研究科）

西山 智明　先生（金沢大学・学際科学実験センター）

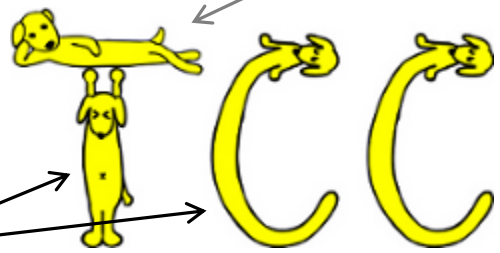
孫 建強　氏（東京大学・大学院農学生命科学研究科・大学院生）

グラント

- 基盤研究(C)（H24-26年度）：「シークエンスに基づく比較トランスクリプトーム解析のためのガイドライン構築」（代表）
- 新学術領域研究（研究領域提案型）（H22年度-）：「非モデル生物におけるゲノム解析法の確立」（分担；研究代表者：西山智明）

挿絵やTCCのロゴなど

（妻の）門田 雅世さま作



（有能な秘書の）三浦 文さま作

次のスライド以降はTCC周辺の詳細な補足説明資料

TbT正規化法(Kadota et al., 2012)の概要

- TCCパッケージに実装している基本コンセプトの原著論文
  - 本来の目的である発現変動遺伝子(DEG)自体がデータ正規化時に悪影響を与えるのでDEG候補を除去して正規化を行うほうがよいこと(DEG Elimination Strategy)を提唱した論文。既存の正規化法は、比較するグループ間でDEG数に偏りがない(unbiased DE)場合にはうまく正規化できるが、**偏りがある場合(biased DE)にはうまく正規化できない**ことを示した。
  - TbT法の実体は、①edgeRパッケージ中のTMM正規化法実行、②baySeqパッケージ中のDEG検出法実行、および③DEG候補を除去した残りのnon-DEG候補のみを用いたTMM正規化法実行、の3ステップを基本とするTMM-baySeq-TMMパイプライン。出力は正規化後の結果（正確には正規化係数）なので、TbT正規化後に任意のDEG検出法を適用することで一連の発現変動解析が終了することになる。例えばTbT正規化法実行後にedgeR中のDEG検出法を適用する一連の手順はTMM-baySeq-TMM-edgeRに相当し、原著論文中ではedgeR/TbTと略記している。論文中ではTbTにした理由を論理的に書いたが、本音は”ToT”に近いものということでTMMとbaySeqを採用。
  - 提案したマルチステップの正規化パイプラインは、第2および第3ステップを繰り返して実行することでより頑健な正規化を実現可能であることも示している。これが図3で説明しているiterative TbT approachに相当するものであり、TMM-$(baySeq-TMM)_n$とも表現できる。例えばiterative TbT正規化法実行後にedgeR中のDEG検出法を適用する一連の手順はTMM-$(baySeq-TMM)_n$-edgeRに相当する。n = 0の場合はTMM-edgeRとなり、これはedgeRパッケージ中のオリジナルの手順と同じである。

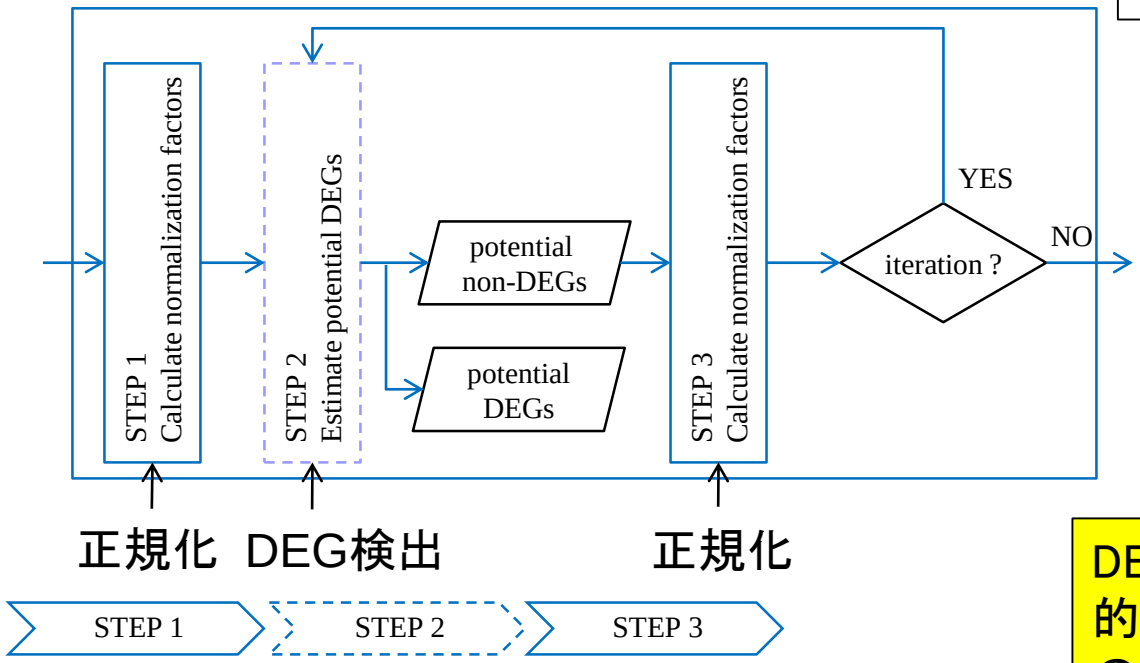
TCC (Sun et al., 2013)の概要

- TbT論文の考え方を一般化し、Rパッケージとしてまとめたという論文
  - TbTはDEG Elimination Strategy (DEGES；でげす)に基づく一つの正規化パイプラインにすぎないこと、第2ステップのbaySeqによるDEG同定ステップが律速であり高速化が課題であったこと、そして各ステップにおいて他の方法が原理的に適用可能であることなどを述べている。
  - 第2ステップのDEG同定法をedgeR中のものに置き換えると、TMM-edgeR-TMMという正規化パイプラインになる。これは、全てedgeRパッケージ中の関数のみで成立するため、DEGES/edgeRと略記している。また、DEGES正規化後にedgeR中のDEG同定法を適用する一連の解析手順は「DEGES/edgeR-edgeR」または「TMM-edgeR-TMM-edgeR」と表記できる。これは実質的にedgeRパッケージ中のオリジナルの解析手順を2回繰り返して行っていることと同義である(ただし、第3ステップのTMMは検出されたDEG候補以外のデータで実行される)。それが、実質的に「TCCは例えばiterative edgeRという理解でよい」と主張する根拠である。
  - TbT論文中で言及したiterative TbTに相当するものは、この論文中ではiterative DEGES (略してiDEGES)と称している。例えば、「iDEGES/edgeR-edgeR」はTMM-(edgeR-TMM)$_n$-edgeRに相当する。n=1は「DEGES/edgeR-edgeR」に相当する。nが2以上の場合がiDEGESに相当するが、nの数を増やしてもその分計算コストがかかる一方で、実質的にn=3程度で頭打ちになることを論文中で示している。それゆえ、iterative DEGESのデフォルトはn=3としている。compcodeR (Soneson, C., *Bioinformatics*, 2014)中でもデフォルトはそうなっている。

DEGESって何デゲス？

～ アラフォー達の略称に関する議論 ～
門田:「DESで行くデス」
西山:「DEGESはいかが？」
門田:「面白くないので却下！」
西山:「左様デゲスか…DEGESって何デゲス？」
門田:「採用！」

- 概念図



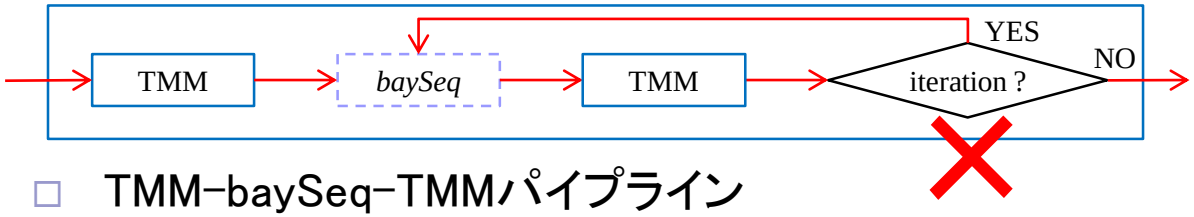
RNA-seqなどから得られるタグカウントデータの正規化をmulti-stepで行う概念の総称

DEG同定を正確に行うのが正規化の目的の一つではあるが、正規化時にDEGの存在自体がDEGとして同定されるのを阻むことがわかった（自爆テロ）。それゆえ、正規化時にDEGの検出を行って、non-DEGのみ利用するのがポイント

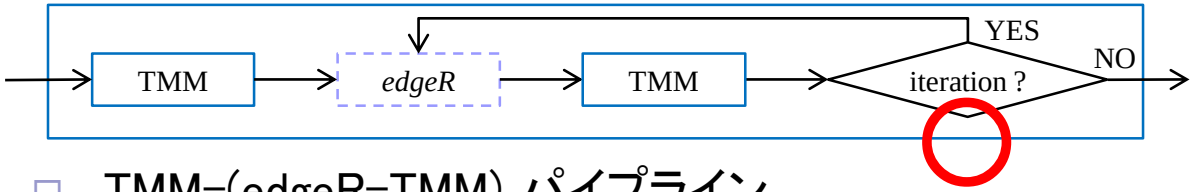
DEGESって何デゲス？

- DEGESのstep1-3で内部的に用いる方法は実用上なんでも?!よい

  - TbT正規化法（Kadota et al., *AMB*, 2012）



    - TMM-baySeq-TMMパイプライン
    - step2でbaySeqパッケージ中のDEG同定法（経験ベイズ）を利用しているため**遅い**…
    - Iterative TbT（step2-3を繰り返してより頑健な正規化係数を得る）は**非現実的**
  - iDEGES/*edgeR*正規化法 (Sun et al., *BMC Bioinformacis*, 2013)

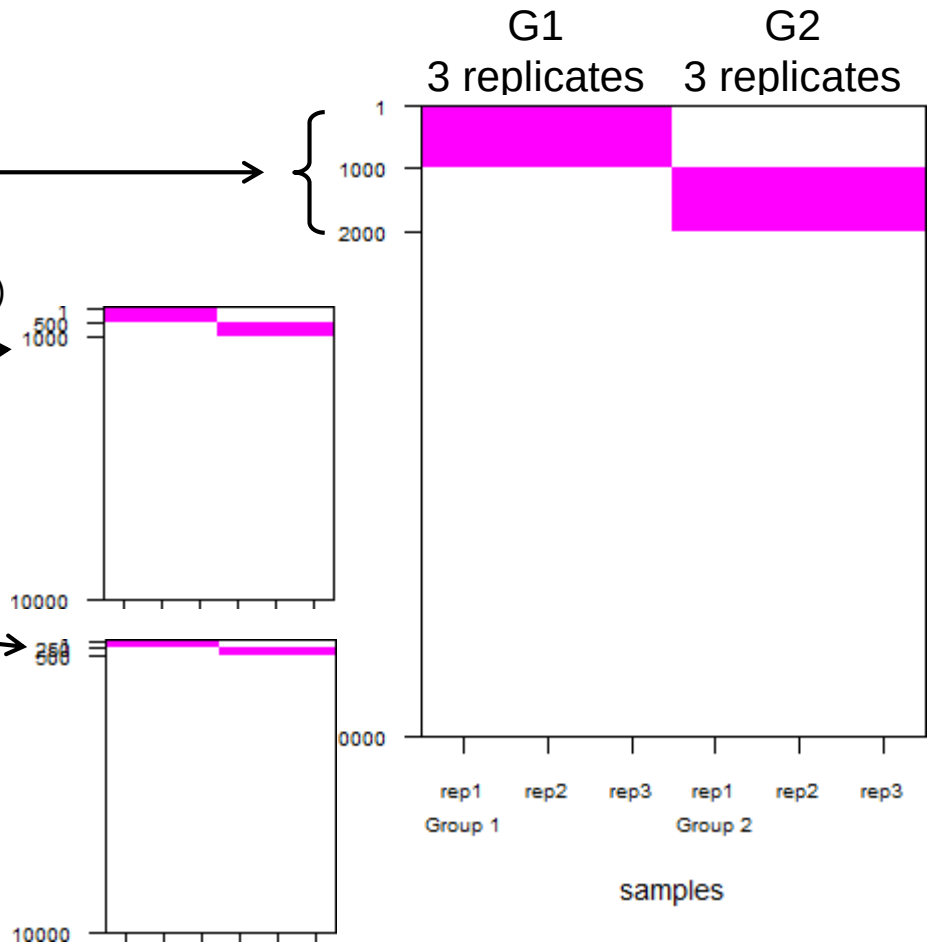


    - TMM-(edgeR-TMM)$_n$パイプライン
    - Step2でedgeRパッケージ中のDEG同定法（exact test）を利用しているため**速い**!
    - DEGESをiterativeに行う頑健なiDEGES（愛デゲス）パイプラインを利用可能

TCCパッケージに実装済み

どういうデータのときに有効デゲスか？

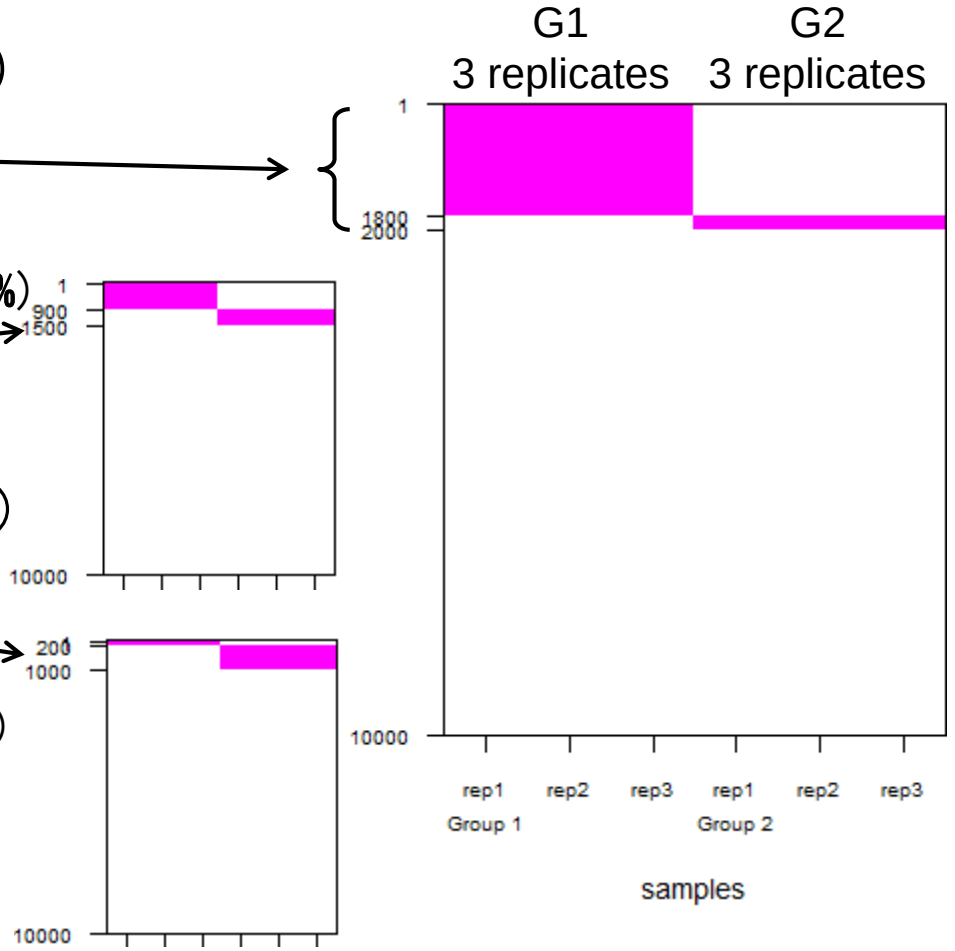
- 仮想データ（10,000 genes × 6 samples）
  - 2,000 DEGs (20%がDEG)
    - Group1 (G1)で高発現：gene1~1000 (**50%**)
    - Group2 (G2)で高発現：gene1001~2000 (**50%**)
  - 1,000 DEGs (10%がDEG)
    - Group1 (G1)で高発現：gene1~500 (**50%**)
    - Group2 (G2)で高発現：gene501~1000 (**50%**)
  - 500 DEGs (5%がDEG)
    - Group1 (G1)で高発現：gene1~250 (**50%**)
    - Group2 (G2)で高発現：gene251~500 (**50%**)

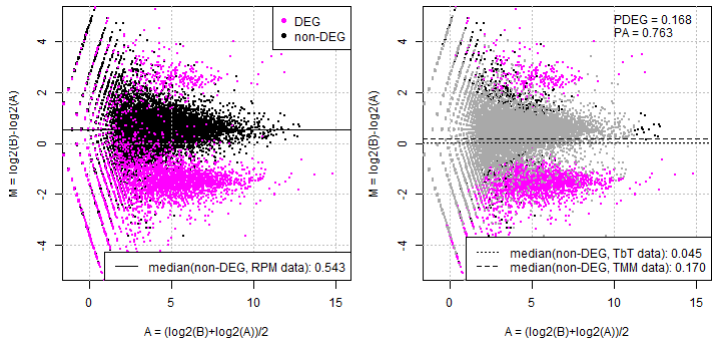


DEG数のGroup間での**偏りがない**場合、TMM正規化法とDEGES系正規化法の**理論上の性能は互角**デゲス。

どういうデータのときに有効デゲスか？

- 仮想データ（10,000 genes × 6 samples）
  - 2,000 DEGs (20%がDEG)
    - Group1 (G1)で高発現：gene1~1800 (**90%**)
    - Group2 (G2)で高発現：gene1801~2000 (**10%**)
  - 1,500 DEGs (15%がDEG)
    - Group1 (G1)で高発現：gene1~900 (**60%**)
    - Group2 (G2)で高発現：gene901~1500 (**40%**)
  - 1,000 DEGs (10%がDEG)
    - Group1 (G1)で高発現：gene1~200 (**20%**)
    - Group2 (G2)で高発現：gene201~1000 (**80%**)



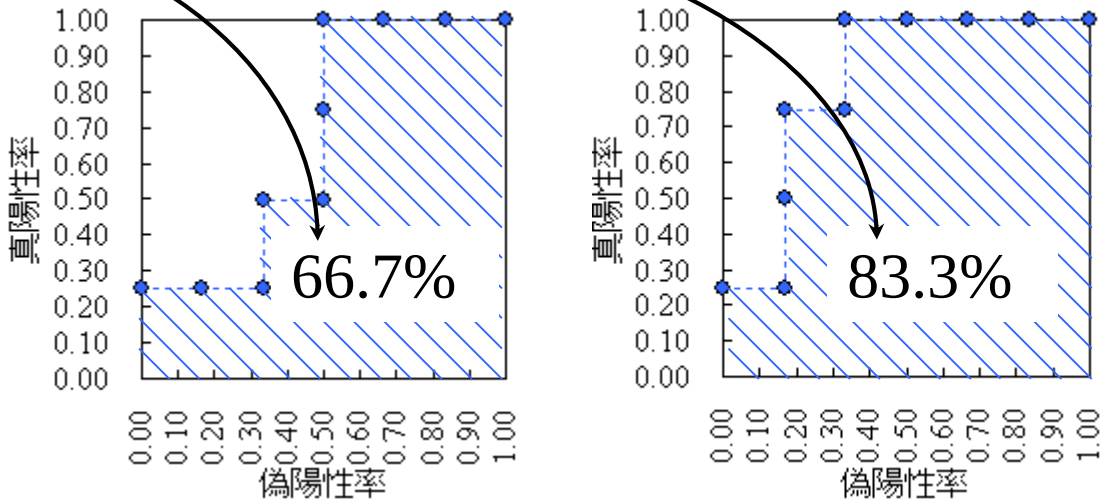


DEGES系正規化法は、DEG数のGroup間での**偏りが大きいほど有効**なんデゲス！

よりよい方法とは？

- その方法を用いて発現変動の度合いでランキングしたときに、真の発現変動遺伝子（DEG）がより上位にランキングされる（感度・特異度高い）

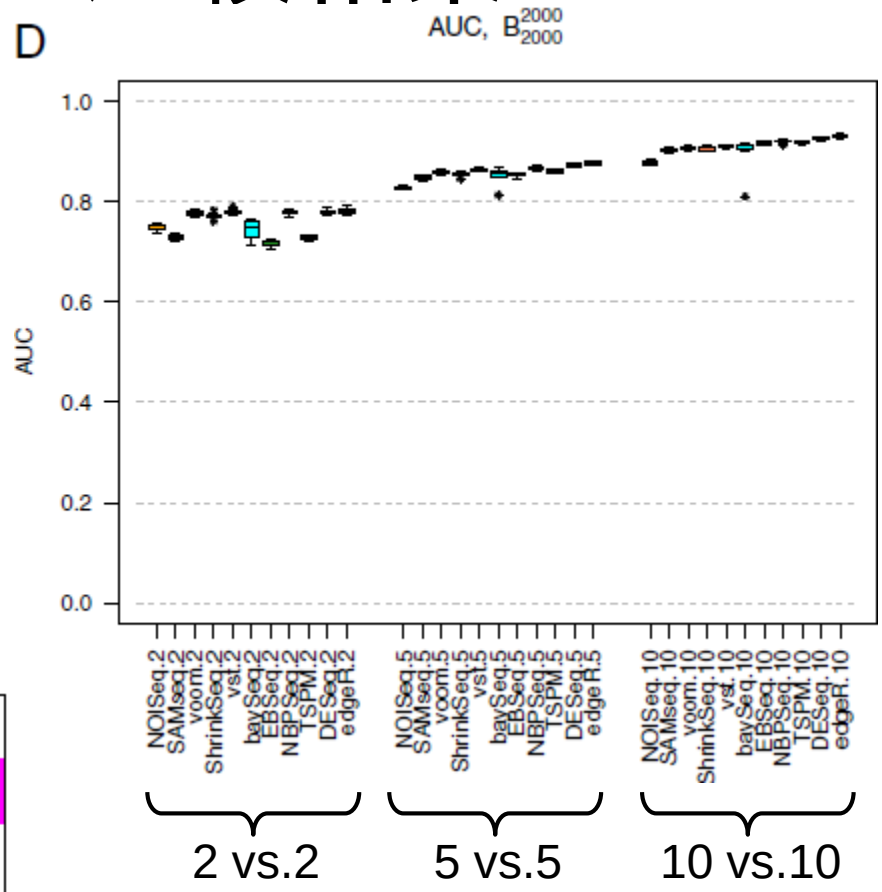
ランキング法				
rank	方法1		方法2	
1	gene8	真	gene8	真
2	gene5	偽	gene5	偽
3	gene4	偽	gene3	真
4	gene3	真	gene2	真
5	gene7	偽	gene7	偽
6	gene1	真	gene1	真
7	gene2	真	gene4	偽
8	gene9	偽	gene9	偽
9	gene10	偽	gene10	偽
10	gene6	偽	gene6	偽

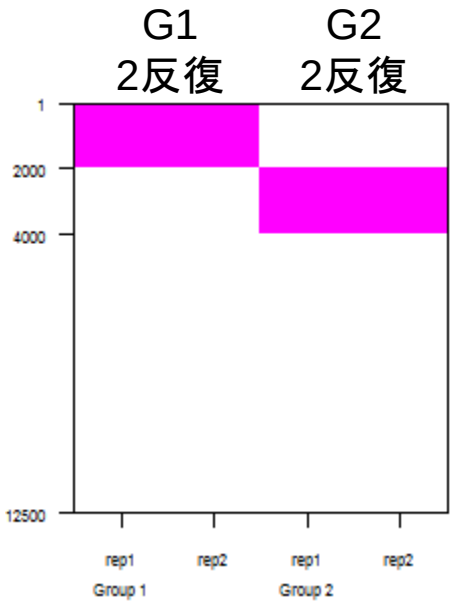


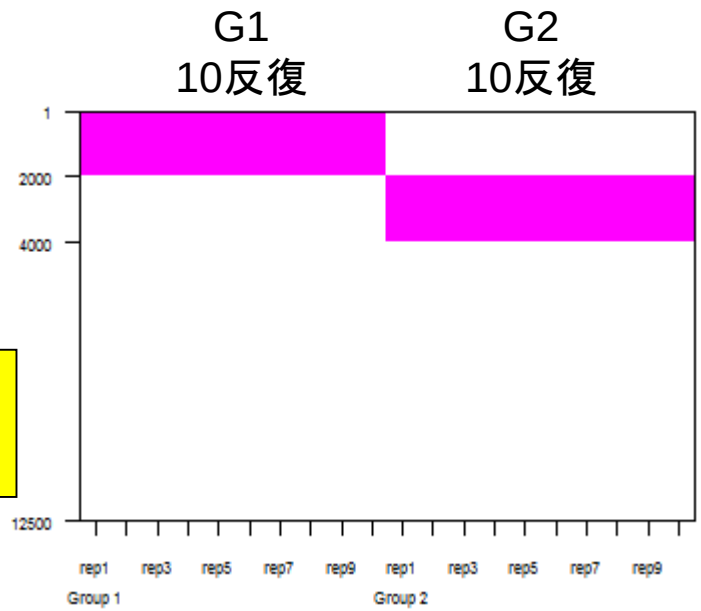
Area Under the ROC Curve（ROC曲線の下部面積：AUC）
バイオインフォマティクス分野でよく用いられる評価基準です

AUC値の比較結果

シミュレーション条件：G1 vs. G2
全遺伝子数：12500
発現変動遺伝子(DEG)数：4000
　G1で高発現：2000
　G2で高発現：2000
unbiased DE situation



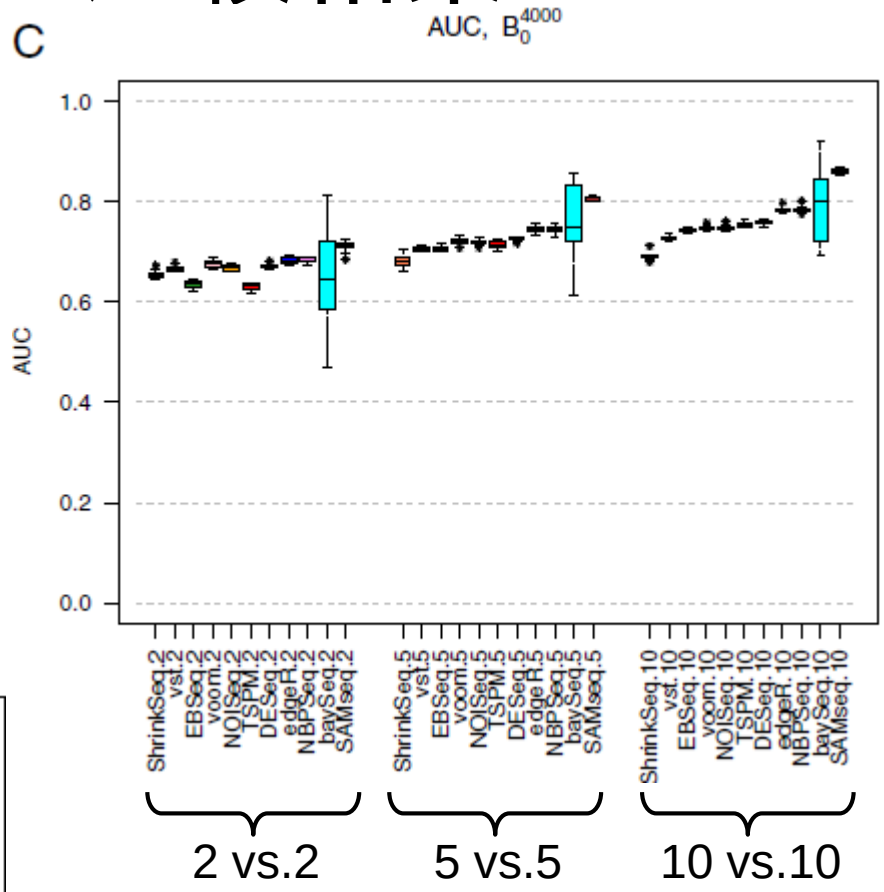


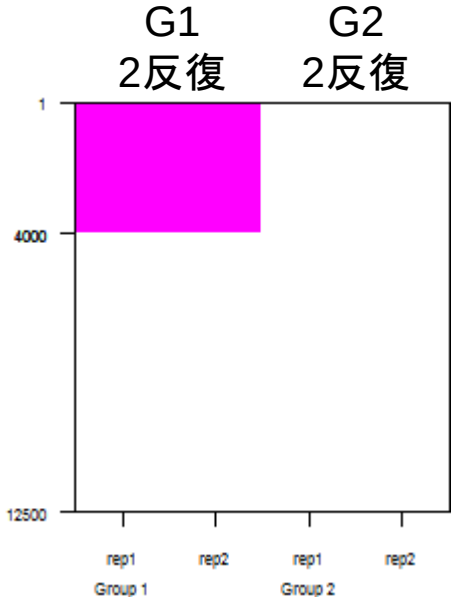


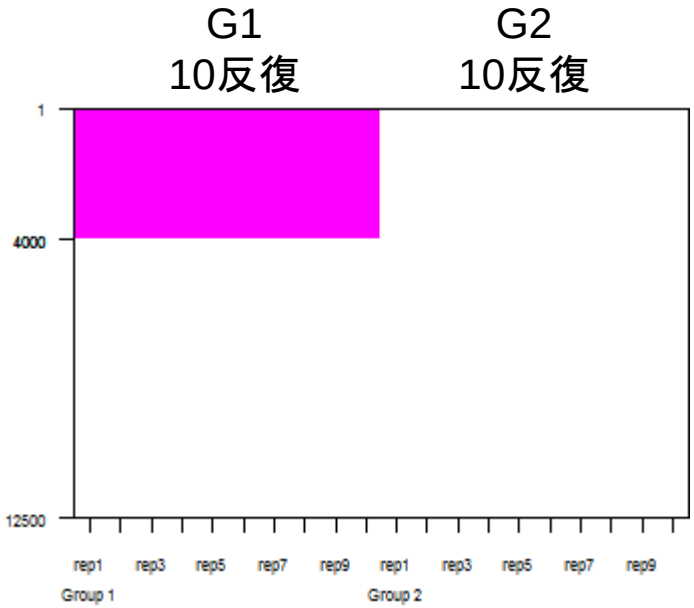
反復実験数を増やすほど精度は上がる
（これが言いたいわけではない…）

AUC値の比較結果

シミュレーション条件：G1 vs. G2
全遺伝子数：12500
発現変動遺伝子(DEG)数：4000
　G1で高発現：4000
　G2で高発現：0
biased DE situation



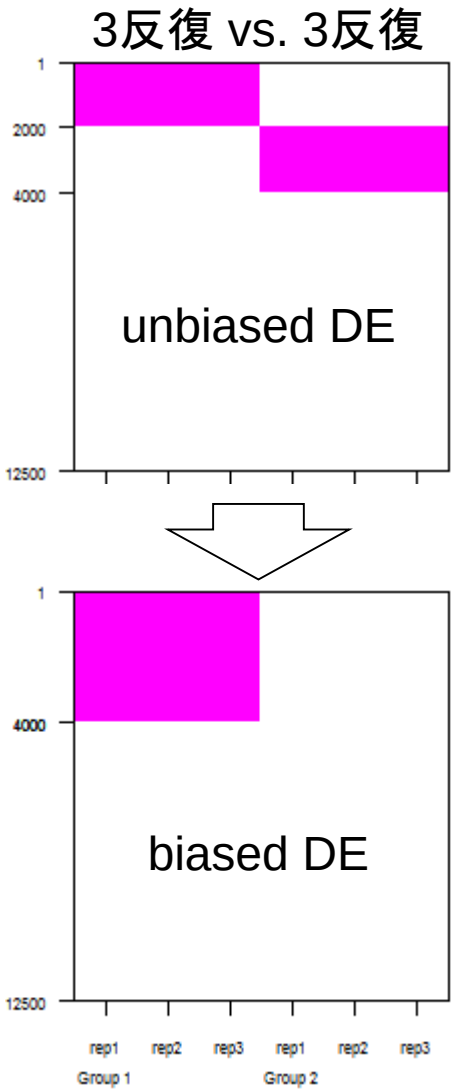




グループ（群）間でDEG数の組成に偏りがあると精度が大幅に低下する

理由：データ正規化法がDEG数の組成に偏りがないことを想定しているため

AUC値の比較結果



	edgeR	SAMseq	*TCC*
unbiased DE	90.84%	87.19%	90.83%
biased DE	82.95%	84.40%	89.92%

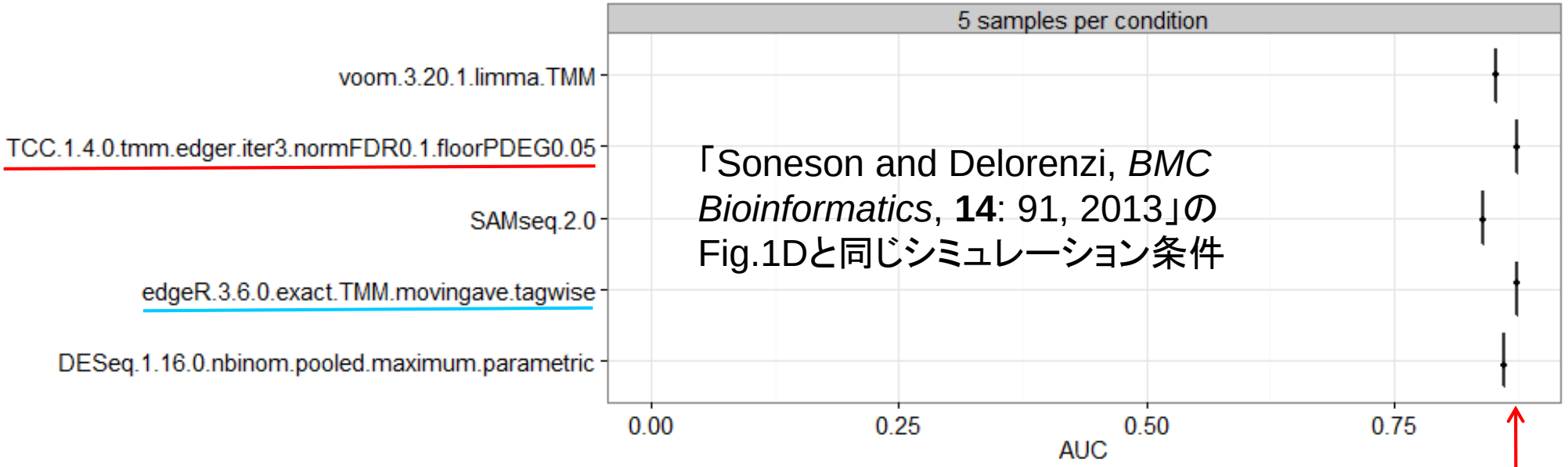
偏りのないデータのときはedgeRがよい
偏りのあるデータのときはSAMseqがよい
→ 偏りの有無に関係なくTCCでよい
ただしこの結果はTCCパッケージ中のシミュレーションデータを用いた評価結果。

compcodeR

- compcodeR：発現変動解析用パッケージ群の性能評価用パッケージ
  - DESeq, DESeq2, SAMseq, edgeR, TCCなどが実装されている
  - TCCパッケージ同様、シミュレーションデータ作成用関数やAUCを含む様々な尺度での性能評価が可能
  - 自作の方法を組み込んで評価することも可能
  - 他分野の似たようなものとしては…
    - Affycompシリーズ：Affymetrixマイクロアレイデータ前処理法の性能評価
    - Assemblathon 2やGAGE：de novo assemblyの性能評価

主に開発者向けのパッケージという位置づけではあるが、オリジナルパッケージ中の生のコードも見られるので、エンドユーザにとってもプログラムを正しい手順で実行できているかの検証用としても有用かも…。

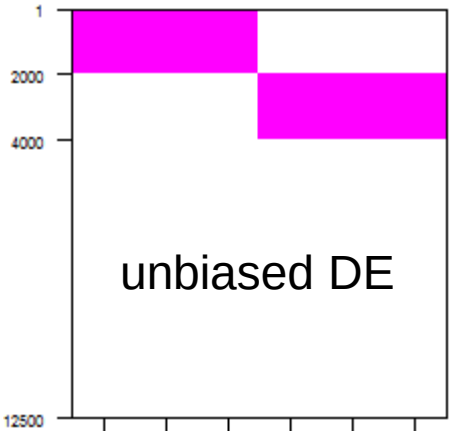
compcodeR (AUCでの評価結果)



「Soneson and Delorenzi, *BMC Bioinformatics*, **14**: 91, 2013」の Fig.1Dと同じシミュレーション条件

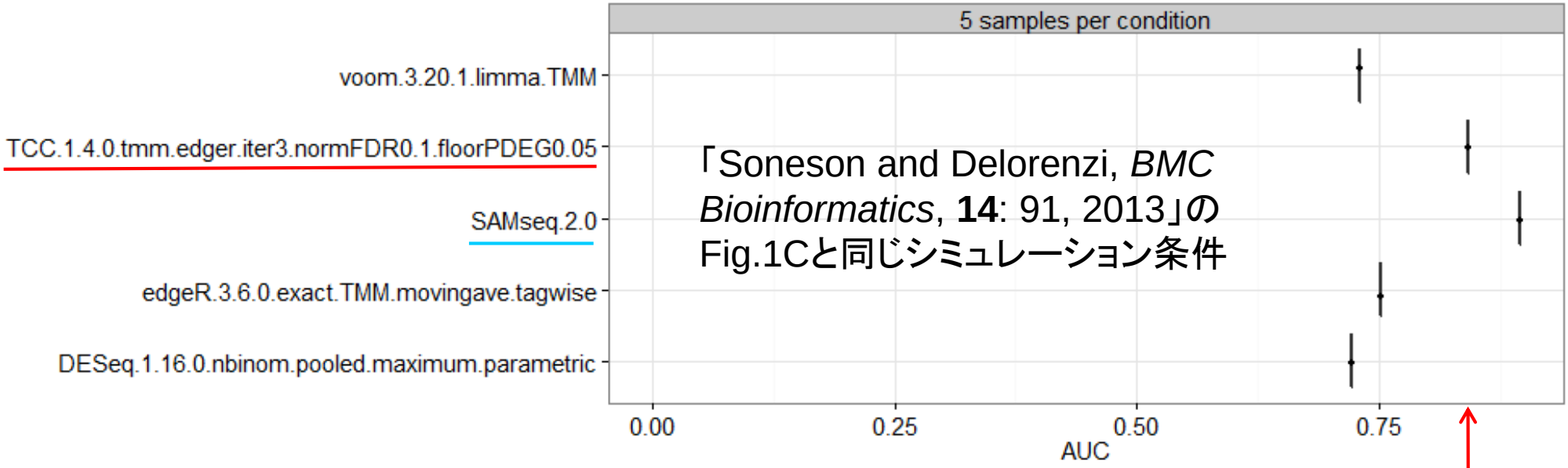
低 ← 感度・特異度 →高

5反復 vs. 5反復



TCCの理論性能通り、compcodeRパッケージで比較してもunbiased DE条件下では、edgeRと同程度のAUC値

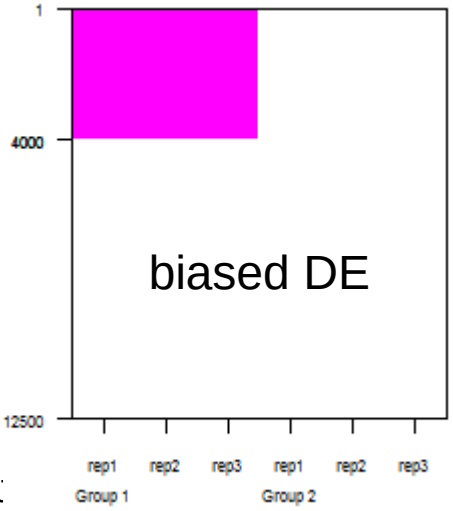
compcodeR (AUCでの評価結果)



「Soneson and Delorenzi, *BMC Bioinformatics*, **14**: 91, 2013」のFig.1Cと同じシミュレーション条件

低 ← 感度・特異度 →高

5反復 vs. 5反復



シミュレーションデータ生成条件などの違いによるのかもしれないが、compcodeRパッケージで比較した場合のbiased DE条件下では、TCCはSAMseqに比べて劣る。しかし、edgeRやDESeqのオリジナルの手順よりも、iterative edgeRやiterative DESeqに相当するTCCのほうが優れていることはcompcodeRパッケージでの性能評価でも示された。